

Network Sampling: From Static to Streaming Graphs

NESREEN K. AHMED, JENNIFER NEVILLE, and RAMANA KOMPELLA, Purdue University

Network sampling is integral to the analysis of social, information, and biological networks. Since many real-world networks are massive in size, continuously evolving, and/or distributed in nature, the network structure is often sampled in order to facilitate study. For these reasons, a more thorough and complete understanding of network sampling is critical to support the field of network science. In this paper, we outline a framework for the general problem of network sampling, by highlighting the different objectives, population and units of interest, and classes of network sampling methods. In addition, we propose a spectrum of computational models for network sampling methods, ranging from the traditionally studied model based on the assumption of a static domain to a more challenging model that is appropriate for streaming domains. We design a family of sampling methods based on the concept of graph induction that generalize across the full spectrum of computational models (from static to streaming) while efficiently preserving many of the topological properties of the input graphs. Furthermore, we demonstrate how traditional static sampling algorithms can be modified for graph streams for each of the three main classes of sampling methods: node, edge, and topology-based sampling. Our experimental results indicate that our proposed family of sampling methods more accurately preserves the underlying properties of the graph for both static and streaming graphs. Finally, we study the impact of network sampling algorithms on the parameter estimation and performance evaluation of relational classification algorithms.

Categories and Subject Descriptors: [Database Management]: Database Applications---Data mining

General Terms: Design, Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Network sampling, statistical network analysis, relational classification

1. INTRODUCTION

Networks arise as a natural representation of data in various domains, ranging from social to biological to information domains. However, the majority of these real-world networks are massive and continuously evolving over time (*i.e.*, streaming). As an example, consider online activity and interaction networks formed from electronic communication (e.g., email, IMs, SMS), social media (e.g., Twitter, blogs, web pages), and content sharing (e.g., Facebook, Flickr, Youtube). These social processes provide a prolific amount of continuous streaming data that is naturally represented as a network where the nodes are people or objects and the edges are the interactions among them (e.g., Facebook users posting 3.2 billion likes and comments every day [AllFacebook.com]). Modeling and analyzing these large dynamic networks have become increasingly important for many applications, such as identifying the behavior and interests of individuals (e.g., viral marketing, online advertising) and investigating how the structure and dynamics of human-formed groups evolve over time.

Unfortunately, many factors make it difficult, if not impossible, to study these networks in their entirety. First and foremost, the sheer size of many networks makes it computationally infeasible to study the entire network. Moreover, some networks are not completely visible to the public (e.g., Facebook) or can only be accessed through crawling (e.g., Web). In other cases, the size of the network may not be as large but the measurements required to observe the underlying network are costly (e.g., experiments in biological networks). Thus, network sampling is at the heart and foundation of our study to understand network structure---since researchers typically need to select a (tractable) subset of the nodes and edges to make inferences about the full network.

From peer-to-peer to social networks, sampling arises across many different settings. For example, sampled networks may be used in simulations and experimentation, to measure performance before deploying new protocols and systems in the field---such as new Internet protocols, social/viral marketing schemes, and/or fraud detection algorithms. In fact, many of the network datasets currently being analyzed as complete networks are themselves samples due to the above limitations in data collection. This means it is critical that researchers understand the impact of sampling methods on the structure of the constructed networks. All of these factors motivate the need for a more refined and complete understanding of *network sampling*. In this paper, we outline a spectrum of compu-

tational models for network sampling and investigate methods of sampling that generalize across this spectrum, going from the simplest and least constrained model focused on sampling from static graphs to the more difficult and most constrained model of sampling from graph streams.

Traditionally, network sampling has been studied in the case of simple static graphs (e.g. [Leskovec and Faloutsos 2006]). These works typically make the simplifying assumption that the graphs are of moderate size and have static structure. Specifically, it is assumed that the graphs fit in the main memory (i.e., algorithms assume the full neighborhood of each node can be explored in a constant time) and many of the intrinsic complexities of realistic networks, such as the time-evolving nature of these systems, are totally ignored. For domains that meet these assumptions, we propose a family of sampling methods based on the concept of graph induction and evaluate our methods against state-of-the-art sampling methods from each of the three classes of network sampling algorithms (node, edge, and topology-based sampling). More importantly, we show that our family of methods preserve the properties of different graphs more accurately than the other sampling methods.

While studying static graphs is indeed important, the assumption that the graph fits in memory is not realistic for many real world domains (e.g., online social networks). When the network is too large to fit in memory, sampling requires random disk accesses that incur large I/O costs. Naturally, this raises the question: how can we sample from these large networks *sequentially*, one edge at a time, while *minimizing* the number of *passes* over the edges? In this context, most of the topology-based sampling procedures such as breadth-first search, random walks, or forest-fire sampling are not appropriate as they require random exploration of a node’s neighbors (which requires many passes over the edges). In contrast, we demonstrate that our sampling methods naturally applies to large graphs, requiring only *two passes* over the edges. Moreover, our proposed sampling algorithm is still able to accurately preserve the properties of the large graph while minimizing the number of passes over the edges (more accurately than alternative algorithms).

Finally, in addition to their massive size, many real-world networks are also likely to be *streaming* over time. A *streaming graph* is a continuous, unbounded, rapid, time-varying stream of *edges* that is clearly too large to fit in memory except for probably short windows of time (e.g., a single day). Streaming graphs occur frequently in the real-world and can be found in many modern online and communication applications such as: Twitter posts, Facebook likes/comments, email communications, network monitoring, sensor networks, among many other applications. Although these domains are quite prevalent, there has been little focus on developing network sampling algorithms that address the complexities of streaming domains. Graph streams differ from static graphs in three main aspects: (i) the massive volume of edges is far too large to fit in the main memory, (ii) the graph structure is not fully observable at any point in time (i.e., only sequential access is feasible, not random access), and (iii) efficient, real-time processing is of critical importance.

The above discussion shows a natural progression of computational models for sampling---from static to streaming. The majority of previous work has focused on sampling from static graphs, which is the simplest and least restrictive problem setup. In this paper, we also focus on the more challenging issues of sampling from disk-resident graphs and graph streams. This leads us to propose a spectrum of computational models for network sampling as shown in Figure 1 where we clearly outline the three computational models for sampling from: (1) static graphs, (2) large graphs, and (3) streaming graphs. This spectrum not only provides insights into the complexity of the computational models (i.e., static vs. streaming), but also the complexity of the algorithms that are designed for each scenario. More complex algorithms are more suitable for the simplest computational model of sampling from static graphs. In contrast, as the complexity of the computational model increases to a streaming scenario, efficient algorithms become necessary. Thus, there is a trade-off between the complexity of the sampling algorithm and the complexity of the computational model (static \rightarrow streaming). A subtle but important consequence is that any algorithm designed to work over graph streams is also applicable in the simpler computational models (i.e., static graphs). However, the converse is not true, algorithms designed for sampling a static graph that can fit in memory,

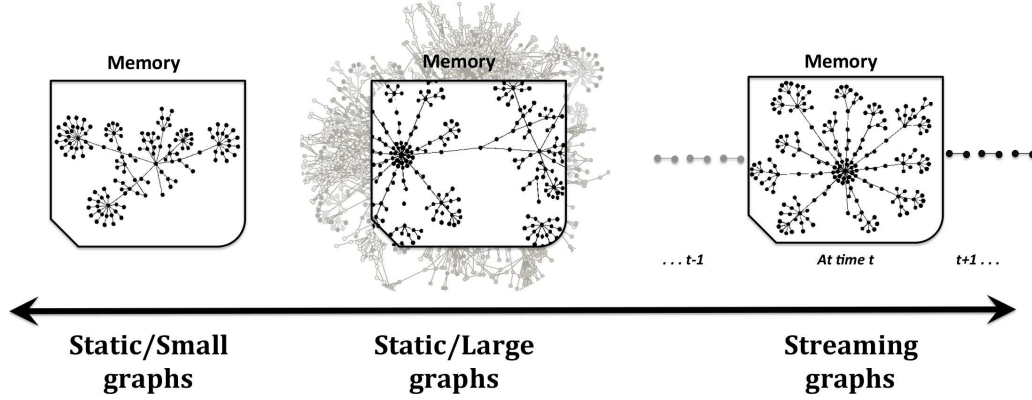


Fig. 1: Spectrum of Computational Models for network sampling: from static to streaming.

may not be generally applicable for graph streams (as they may require an intractable number of passes to implement).

Within the spectrum of computational models for network sampling, we formally discuss in Section 2 the various objectives of network sampling (e.g., sampling to estimate network parameters). We provide insights on how conventional objectives in static domains can be naturally adapted to the more challenging scenario of streaming graphs. In Sections 5 and 6, we primarily focus on the task of representative subgraph sampling from both static and streaming graphs. As an example problem definition, we consider the case of sampling representative subgraphs from graph streams. Formally, the input is assumed to be a graph $G = (V, E)$, presented as a stream of edges E in no particular order. Then, the goal is to sample a subgraph $G_s = (V_s, E_s)$ with a subset of the nodes ($V_s \subset V$) and/or edges ($E_s \subset E$) from the population graph stream G . The objective is to ensure that G_s is a *representative* subgraph that matches many of the topological properties of G . In addition to the *sample representativeness* requirement, a graph stream sampling algorithm is also required to be *efficient*---and needs to decide whether to include an edge $e \in E$ in the sample or not as the edge is *streamed* in. The sampling algorithm may maintain a state $|\Psi|$ and consult the state to determine whether to sample the edge or not, but the total storage associated with the algorithm is preferred to be of the order the size of the output sampled subgraph G_s , i.e., $|\Psi| = O(|G_s|)$.

In this paper, we formalize the problem of sampling from graph streams. We show how to extend traditional sampling algorithms from each of the three classes of sampling methods (node, edge, and topology-based sampling) for use on graph streams. For edge sampling from graph streams, we use the approach in [Aggarwal et al. 2011]. Furthermore, we propose our graph stream sampling method (from the family of methods based on the concept of graph induction), which is space-efficient (uses space in the order of the sampled subgraph) and runs in a *single pass* over the edges. Our family of sampling methods based on the concept of graph induction generalize across the full spectrum of computational models (from static to streaming) while efficiently preserving many of the graph properties for streaming and static graphs. In addition, our proposed family of sampling methods offers a good balance between algorithm complexity and sample representativeness while remaining general enough for any computational model. Notably, our family of algorithms, while less complex, preserve the properties of *static* graphs even better than the more complex algorithms that do not generalize to the streaming model, over a broad set of real-world networks.

In conclusion, we summarize the contributions of this paper as follows:

- A detailed framework outlining the general problem of network sampling, highlighting the different goals, population and units, and classes of network sampling methods (see Section 2).

- A further elaboration of the above framework to include a spectrum of computational models within which to design network sampling methods (*i.e.*, going from static to streaming graphs) (see Section 3).
- Introduction of a family of sampling methods based on the concept of *graph induction* that has the following properties (see Sections 5 and 6):
 - (1) Preserve many key graph characteristics more accurately than alternative state-of-the-art algorithms.
 - (2) Generalize to a streaming computational model with a minimum storage space (*i.e.*, space complexity is on the order of the sample size).
 - (3) Run efficiently in a few number of passes over the edges (*i.e.*, runtime complexity is on the order of the number of edges).
- Systematic investigation of the three classes of static sampling methods on a variety of datasets and extension of these algorithms for application to graph streams (see Sections 5 and 6).
- Empirical evaluation that shows our sampling methods are applicable to large graphs that don't fit in the main memory (see Sections 5 and 6).
- Further task-based evaluation of sampling algorithm performance in the context of relational classification. This investigation illustrates the impact of network sampling on the parameter estimation and evaluation of classification algorithms overlaid on the sampled networks (see Section 7).

2. FOUNDATIONS OF NETWORK SAMPLING

In the context of statistical data analysis, a number of issues arise and need to be considered carefully before collecting data and making inferences based on them. At first, we need to identify the relevant *population* to be studied. Then, if sampling is necessary then we need to decide how to sample from that population. Generally, the term *population* is defined as the full set of representative units that one wishes to study (*e.g.*, units may be individuals in a particular city). In some instances, the population may be relatively small and bounded, and is therefore easy to study in its entirety (*i.e.*, without sampling). For instance, it is relatively easy to study the set of graduate students in an academic department. Conversely, in other situations the population may be large, unbounded, or difficult and/or costly to access in its entirety. For instance, the complete set of Facebook users. In this case, a sample of units should be collected and characteristics of the population can be estimated from the sampled units.

Network sampling is of interest to a variety of researchers from many distinct fields (*e.g.* statistics, social science, databases, data mining, machine learning) due to the range of complex datasets that can be represented as graphs. While each area may investigate different types of networks, they all have focused primarily on *how* to sample.

For example, in social science, snowball sampling is used extensively to run survey sampling in populations that are difficult-to-access (*e.g.*, sampling the set of drug users in a city) [Watters and Biernacki 1989]. Similarly, in Internet topology measurements, breadth first search is used to *crawl* distributed, large-scale Online social networks (*e.g.* Facebook) [Mislove et al. 2007]. Moreover, in structured data mining and machine learning, the focus has been on developing algorithms to sample small(er) subgraphs from a single large network [Leskovec and Faloutsos 2006]. These sampled subgraphs are further used to learn models (*e.g.*, relational classification models [Friedman et al. 1999]), evaluate and compare the performance of algorithms (*e.g.*, different classification methods [Rossi and Neville 2012]), and study complex network processes (*e.g.*, information diffusion [Bakshy et al. 2012]). We provide a detailed discussion of the related work in Section 4.

While this large body of research has developed methods to sample from networks, much of the work is problem-specific and there has been less work focused on developing a broader foundation for network sampling. More specifically, it is often not clear *when* and *why* particularly sampling methods are appropriate. This is because the goals and population are often not explicitly defined or stated up front, which makes it difficult to evaluate the quality of the recovered samples for

other applications. One of the primary aims of this paper is to define and discuss the foundations of network sampling more explicitly, such as: objectives/goals, population of interest, units, classes of sampling algorithms (*i.e.*, node, edge, and topology-based), and techniques to evaluate a sample (*e.g.*, network statistics and distance metrics). In this section, we will outline a solid methodological framework for network sampling. The framework will facilitate the comparison of various network sampling algorithms, and help to understand their relative strengths and weaknesses with respect to particular sampling goals.

2.1. Notation

Formally, we consider an input network represented as a graph $G = (V, E)$ with the node set $V = \{v_1, v_2, \dots, v_N\}$ and edge set $E = \{e_1, e_2, \dots, e_M\}$, such that $N = |V|$ is the number of nodes, and $M = |E|$ is the number of edges. We denote $\eta(\cdot)$ as any topological graph property. Therefore, $\eta(G)$ could be a *point statistic* (*e.g.*, average degree of nodes in V) or a *distribution* (*e.g.*, degree distribution of V in G).

Further, we define $\Lambda = \{a_1, a_2, \dots, a_k\}$ as the set of k attributes associated with the nodes describing their properties. Each node $v_i \in V$ is associated with an attribute vector $[a_1(v_i), a_2(v_i), \dots, a_k(v_i)]$ where $a_j(v_i)$ is the j^{th} attribute value of node v_i . For instance, in a Facebook network where nodes represent users and edges represent friendships, the node attributes may include age, political view, and relationship status of the user.

Similarly, we denote $\beta = \{b_1, b_2, \dots, b_l\}$ as the set of l attributes associated with the edges describing their properties. Each edge $e_{ij} = (v_i, v_j) \in E$ is associated with an attribute vector $[b_1(e_{ij}), b_2(e_{ij}), \dots, b_l(e_{ij})]$. In the Facebook example, edge attributes may include relationship type (*e.g.*, friends, married), relationship strength, and type of communication (*e.g.*, wall post, picture tagging).

Now, we define the network sampling process. Let σ be any sampling algorithm that selects a random sample S from G (*i.e.*, $S = \sigma(G)$). The sampled set S could be a subset of the nodes ($S = V_s \subset V$), or edges ($S = E_s \subset E$), or a subgraph ($S = (V_s, E_s)$ where $V_s \subset V$ and $E_s \subset E$). The size of the sample S is defined relative to the graph size with respect to a sampling fraction ϕ ($0 \leq \phi \leq 1$). In most cases the sample size is defined as a fraction of the nodes in the input graph, *e.g.*, $|S| = \phi \cdot |V|$. But in some cases, the sample size is defined relative to the number of edges ($|S| = \phi \cdot |E|$).

2.2. Goals, Units, and Population

While the explicit aim of many network sampling algorithms is to select a smaller subgraph from G , there are often other more implicit goals of the process that are left unstated. Here, we formally outline a range of possible goals for network sampling:

(1) ESTIMATE NETWORK PARAMETERS

Used to select a subset S of the nodes (or edges) from G , to estimate properties of G . Thus, S is a good sample of G if,

$$\eta(S) \approx \eta(G)$$

For example, let $S = V_s \subset V$ be the subset of sampled nodes, we can estimate the average degree of nodes G using S as

$$\hat{deg}_{avg} = \frac{1}{|S|} \sum_{v_i \in S} deg(v_i \in G)$$

where $deg(v_i \in G)$ is the degree of node v_i as it appears in G , and a direct application of statistical estimators helps to correct the sampling bias of \hat{deg}_{avg} [Hansen and Hurwitz 1943].

(2) SAMPLE A REPRESENTATIVE SUBGRAPH

Used to select a small subgraph $S = G_s = (V_s, E_s)$ from G , such that S preserves some

topological properties of G . Let η_A be a set of topological properties, then S is a good sample of G if,

$$\eta_A(S) \approx \eta_A(G)$$

Generally, the subgraph representativeness is evaluated by picking a set of graph topological properties that are important for a wide range of applications. This ensures that the sample subgraph S can be used instead of G for testing algorithms, systems, and/or models in an application. For example, [Leskovec and Faloutsos 2006] uses topological properties like degree, clustering, and eigenvalues to evaluate the samples.

(3) ESTIMATE NODE ATTRIBUTES

Used to select a subset $S = V_s$ of the nodes from G to study node attributes. Let f_a be a function involving node attribute a , then, $S \subset V$ is a good sample of V if,

$$f_a(S) \approx f_a(V)$$

For example, if a represents the age of users, we can estimate the average age in G using S as

$$\hat{a}_{avg} = \frac{1}{|S|} \sum_{v_i \in S} a(v_i)$$

Similar to goal 1, statistical estimators can be used to correct the bias.

(4) ESTIMATE EDGE ATTRIBUTES

Used to select a subset $S = E_s$ of the edges from G to study edge attributes. Let f_b be a function involving edge attribute b , then, $S \subset E$ is a good sample of E if,

$$f_b(S) \approx f_b(E)$$

For example, if b represents the relationship type of friends (e.g., married, coworkers), we can estimate the proportion of married relationships in G using S as

$$\hat{p}_{married} = \frac{1}{|S|} \sum_{e_{ij} \in S} 1_{(b(e_{ij})=married)}$$

Clearly, the first two goals (1 and 2) focus on characteristics of entire networks, while the last two goals (3 and 4) focus on characteristics of nodes or edges in isolation. Therefore, these goals maybe difficult to satisfy simultaneously---i.e., if the sampled data enable accurate study of one, it may not allow accurate study of the others. For instance, a representative subgraph sample could be a biased estimate of global graph parameters (e.g., density).

Once the goal is outlined, the population of interest can be defined relative to the goal. In many cases, the definition of the population may be obvious. The main challenge is then to select a representative subset of units in the population in order to make the study cost efficient and feasible. Other times, the population may be less tangible and difficult to define. For example, if one wishes to study the characteristics of a system or *process*, there is not a clearly defined set of items to study. Instead, one is often interested in the overall behavior of the system. In this case, the population can be defined as the set of possible outcomes from the system (e.g., measurements over all settings) and these *units* should be sampled according to their underlying probability distribution.

In the first two goals outlined above, the objective of study is an entire network (either for structure or parameter estimation). In goal 1, if the objective is to estimate local properties from the nodes (e.g. degree distribution of G), then the elementary units are the nodes, and then the population would be the set of all nodes V in G . However, if the objective is to estimate global properties (e.g. diameter of G), then the elementary units correspond to subgraphs (any $G_s \subset G$) rather than nodes and the population should be defined as the set of subgraphs of a particular size that could

be drawn from G . In goal 2, the objective is to select a subgraph G_s , thus the elementary units correspond to subgraphs, rather than nodes or edges (goal 3 and 4). As such, the population should also be defined as the set of subgraphs of a particular size that could be drawn from G .

2.3. Classes of Sampling Methods

Once the population has been defined, a sampling algorithm σ must be chosen to sample from G . Sampling algorithms can be categorized as node, edge, and topology-based sampling, based on whether nodes or edges are locally selected from G (node and edge-based sampling) or if the selection of nodes and edges depends more on the existing topology of G (topology-based sampling). Graph sampling algorithms have two basic steps:

- (1) *Node selection*: used to sample a subset of nodes $S = V_s$ from G , (*i.e.*, $V_s \subset V$).
- (2) *Edge selection*: used to sample a subset of edges $S = E_s$ from G , (*i.e.*, $E_s \subset E$)

When the objective is to sample only nodes or edges (*i.e.*, goals 1, 3, 4), then either step 1 or step 2 is used to form the sample S . When the objective is to sample a subgraph G_s from G (*i.e.*, goal 2), then both step 1 and 2 from above are used to form S , (*i.e.*, $S = (V_s, E_s)$). In this case, the edge selection is often conditioned on the selected node set in order to form an *induced subgraph* by sampling a subset of the edges incident to V_s (*i.e.* $E_s = \{e_{ij} = (v_i, v_j), e_{ij} \in E | v_i, v_j \in V_s\}$). We distinguish between two different approaches to graph induction---*total* and *partial* graph induction--which differ by whether *all* or *some* of the edges incident on V_s are selected. The resulting sampled graphs are referred to as the *induced subgraph* and *partially induced subgraph* respectively.

While the discussion of the algorithms in the next sections focuses more on sampling a subgraph G_s from G , they can easily generalize to sampling only nodes or edges.

Node sampling (NS). In classic node sampling, nodes are chosen independently and uniformly at random from G for inclusion in the sampled graph G_s . For a target fraction ϕ of nodes required, each node is simply sampled with a probability of ϕ . Once the nodes are selected for V_s , the sampled subgraph is constructed to be the *induced subgraph* over the nodes V_s , *i.e.*, all edges among the $V_s \in G$ are added to E_s . While node sampling is intuitive and relatively straightforward, the work in [Stumpf et al. 2005] shows that it does not accurately capture properties of graphs with power-law degree distributions. Similarly, [Lee et al. 2006] shows that although node sampling appears to capture nodes of different degrees well, due to its inclusion of all edges for a chosen node set only, the original level of connectivity is not likely to be preserved.

Edge sampling (ES). In classic edge sampling, edges are chosen independently and uniformly at random from G for inclusion in the sampled graph G_s . Since edge sampling focuses on the selection of edges rather than nodes to populate the sample, the node set is constructed by including both incident nodes in V_s when a particular edge is sampled (and added to E_s). The resulting subgraph is partially induced, which means no extra edges are added over and above those that were chosen during the random edge selection process. Unfortunately, ES fails to preserve many desired graph properties. Due to the independent sampling of edges, it does not preserve clustering and connectivity. It is however more likely to capture path lengths, due to its bias towards high degree nodes and the inclusion of both end points of selected edges.

Topology-based sampling. Due to the known limitations of NS ([Stumpf et al. 2005; Lee et al. 2006]) and ES (bias toward high degree nodes), researchers have also considered many other topology-based sampling methods, which use breadth-first search (*i.e.* sampling without replacement) or random walks (*i.e.* sampling with replacement) over the graph to construct a sample.

One example is snowball sampling, which adds nodes and edges using breadth-first search (but with only a fraction of neighbors explored) from a randomly selected seed node. Snowball sampling accurately maintains the network connectivity within the snowball, however it suffers from *boundary bias* in that many peripheral nodes (*i.e.*, those sampled on the last round) will be missing a large number of neighbors [Lee et al. 2006].

Table I: Description of Network Statistics

Network Statistic	Description
DEGREE DIST.	Distribution of nodes degrees in the network
PATH LENGTH DIST.	Distribution of all shortest paths
CLUSTERING COEFFICIENT DIST.	Distribution of local clustering per node
K-CORE DIST.	Distribution of sizes of the largest subgraphs where nodes have at least k interconnections
EIGENVALUES	Distribution of the eigenvalues of the network adjacency matrix vs. their rank
NETWORK VALUES	Distribution of eigenvector components of the largest eigenvalue of the network adjacency matrix vs. their rank

Another example is the Forest Fire Sampling (FFS) method [Leskovec and Faloutsos 2006], which uses *partial* breadth-first search where only a fraction of neighbors are followed for each node. The algorithm starts by picking a node uniformly at random and adding it to the sample. It then "burns" a random proportion of its outgoing links, and adds those edges, along with the incident nodes, to the sample. The fraction is determined by sampling from a geometric distribution with mean $p_f/(1 - p_f)$. The authors recommend setting $p_f = 0.7$, which results in an average burn of 2.33 edges per node. The process is repeated recursively for each burned neighbor until no new node is selected, then a new random node is chosen to continue the process until the desired sample size is obtained. Also, there are other examples such as respondent-driven sampling [Heckathorn 1997] and expansion sampling [Maiya and Berger-Wolf 2010], we give more details in Section 4.

In general, such topology-based sampling approaches form the sampled graph out of the explored nodes and edges, and usually perform better than simple algorithms such as NS and ES.

2.4. Evaluation of Sampling Methods

When the goal is to study the entire input network---either for measuring the quality of parameter estimates (goal 1), or measuring the representativeness of the sampled subgraph structure (goal 2)--the accuracy of network sampling methods is often evaluated by comparing network statistics (*e.g.*, degree). We first define a suite of common network statistics and then discuss how they can be used more quantitatively to compare sampling methods.

Network Statistics. The commonly considered network statistics can be compared along two dimensions: *local* vs. *global* statistics, and *point* statistic vs. *distribution*. A local statistic is used to describe a characteristic of a local graph element (*e.g.*, node, edge, subgraph). For example, node degree and node clustering coefficient. On the other hand, a global statistic is used to describe a characteristic of the entire graph. For example, global clustering coefficient and graph diameter. Similarly, there is also the distinction between point statistics and distributions. A point-statistic is a single value statistic (*e.g.*, diameter) while a distribution is a multi-valued statistic (*e.g.*, distribution of path length for all pairs of nodes). Clearly, a range of network statistics are important to understand the full graph structure.

In this work, we focus on the goal of sampling a representative subgraph G_s from G , by using distributions of network characteristics calculated on the level of nodes, edges, sets of nodes or edges, and subgraphs. Table I provides a summary for the six network statistics we use and we formally define the statistics below:

- (1) *Degree distribution*: The fraction of nodes with degree k , for all $k > 0$

$$p_k = \frac{|\{v \in V | \deg(v) = k\}|}{N}$$

Degree distribution has been widely studied by many researchers to understand the connectivity in graphs. Many real-world networks were shown to have a power-law degree distribution, for example in the Web [Kleinberg et al. 1999], citation graphs [Redner 1998], and online social networks [Chakrabarti et al. 2004].

- (2) *Path length distribution*: Also known as the *hop* distribution and denotes the fraction of pairs $(u, v) \in V$ with a shortest-path distance ($\text{dist}(u, v)$) of h , for all $h > 0$

$$p_h = \frac{|\{(u, v) \in V | \text{dist}(u, v) = h\}|}{N^2}$$

The path length distribution is essential to know how the number of paths between nodes expands as a function of distance (*i.e.*, number of hops).

- (3) *Clustering coefficient distribution*: The fraction of nodes with clustering coefficient ($\text{cc}(v)$) c , for all $0 \leq c \leq 1$

$$p_c = \frac{|\{v \in V' | \text{cc}(v) = c\}|}{|V'|}, \text{ where } V' = \{v \in V | \deg(v) > 1\}$$

Here the clustering coefficient of a node v is calculated as the number of triangles centered on v divided by the number of pairs of neighbors of v (*e.g.*, the proportion of v 's neighbor that are linked). In social networks and many other real networks, nodes tend to cluster. Thus, the clustering coefficient is an important measure to capture the transitivity of the graph [Watts and Strogatz 1998].

- (4) *K-core distribution*: The fraction of nodes in graph G participating in a k -core of order k . The k -core of G is the largest induced subgraph with minimum degree k . Formally, let $U \subseteq V$, and $G_{[U]} = (U, E')$ where $E' = \{e_{u,v} \in E | u, v \in U\}$. Then $G_{[U]}$ is a k -core of order k if $\forall v \in U \deg_{G_{[U]}}(v) \geq k$.

Studying k -cores is an essential part of social network analysis as they demonstrate the connectivity and community structure of the graph [Carmi et al. 2007; Alvarez-Hamelin et al. 2005; Kumar et al. 2010]. We denote the *maximum core number* as the maximum value of k in the k -core distribution. The *maximum core number* can be used as a lower bound on the degree of the nodes that participate in the largest induced subgraph of G . Also, the core sizes can be used to demonstrate the localized density of subgraphs in G [Seshadhri et al. 2011].

- (5) *Eigenvalues*: The set of real eigenvalues $\lambda_1 \geq \dots \geq \lambda_N$ of the corresponding adjacency matrix A of G . Since *eigenvalues* are the basis of spectral graph analysis, we compare the largest 25 eigenvalues of the sampled graphs to their real counterparts.
- (6) *Network values*: The distribution of the eigenvector components associated with the largest eigenvalue λ_{\max} . We compare the largest 100 network values of the sampled graphs to their real counterparts.

Next, we describe the use of these statistics for comparing sampling methods quantitatively.

Distance Measures for Quantitatively Comparing Sampling Methods.

The goal is to select a *representative* sample that minimizes the distance between the property in G and the property in G_s : $\text{dist}[\eta(G), \eta(G_s)]$. When the goal is to provide estimates of global network parameters (*e.g.*, average degree), then $\eta(\cdot)$ may measure *point statistics*. However, when the goal is to provide a representative subgraph sample, then $\eta(\cdot)$ may measure *distributions* of network properties (*e.g.*, degree distribution). These distributions reflects how the graph structure is distributed across nodes and edges.

The *dist* function could be typically any distance measure (e.g., absolute difference). In this paper, since we focus on using distributions to characterize graph structure, we use four different distributional distance measures for evaluation.

- (1) *Kolmogorov-Smirnov (KS) statistic*: Used to assess the distance between two cumulative distribution functions (CDF). The KS-statistic is a widely used measure of the agreement between two distributions, including in [Leskovec and Faloutsos 2006] where it is used to illustrate the accuracy of FFS. It is computed as the maximum vertical distance between the two distributions, where x represents the range of the random variable and F_1 and F_2 represent two CDFs:

$$KS(F_1, F_2) = \max_x |F_1(x) - F_2(x)|$$

- (2) *Skew divergence (SD)*: Used to assess the difference between two probability density functions (PDF) [Lee 2001]. Skew divergence is used to measure the Kullback-Leibler (KL) divergence between two PDFs P_1 and P_2 that do not have continuous support over the full range of values (e.g., skewed degree). KL measures the average number of extra bits required to represent samples from the original distribution when using the sampled distribution. However, since KL divergence is not defined for distributions with different areas of support, skew divergence *smooths* the two PDFs before computing the KL divergence:

$$SD(P_1, P_2, \alpha) = KL[\alpha P_1 + (1 - \alpha)P_2 \parallel \alpha P_2 + (1 - \alpha)P_1]$$

The results shown in [Lee 2001] indicate that using SD yields better results than other methods to approximate KL divergence on non-smoothed distributions. In this paper, as in [Lee 2001], we use $\alpha = 0.99$.

- (3) *Normalized L_1 distance*: In some cases, for evaluation we will need to measure the distance between two positive m -dimensional real vectors p and q such that p is the true vector and q is the estimated vector. For example, to compute the distance between two vectors of eigenvalues. In this case, we use the normalized L_1 distance:

$$L_1(p, q) = \frac{1}{m} \sum_{i=1}^m \frac{|p_i - q_i|}{p_i}$$

- (4) *Normalized L_2 distance*: In other cases, when the vector components are fractions (less than one), we use the normalized euclidean distance L_2 distance (e.g., to compute the distance between two vectors of network values):

$$L_2(p, q) = \frac{\|p - q\|}{\|p\|}$$

3. MODELS OF COMPUTATION

In this section, we discuss the different models of computation that can be used to implement network sampling methods. At first, let us assume the network $G = (V, E)$ is given (e.g., stored on a large storage device). Then, the goal is to select a sample S from G .

Traditionally, network sampling has been explored in the context of a *static model of computation*. This simple model makes the fundamental assumption that it is easy and fast (i.e., constant time) to randomly access any location of the graph G . For example, random access may be used to query the entire set of nodes V or to query the neighbors $\mathcal{N}(v_i)$ of a particular node v_i (where $\mathcal{N}(v_i) = \{v_j \in V | e_{ij} = (v_i, v_j) \in E\}$). However, random accesses on disks are much slower than random accesses in main memory. A key disadvantage of the static model of computation is that it does not differentiate between a graph that can fit entirely in the main memory and a graph that cannot. Conversely, the primary advantage of the static model is that, since it is the natural extension of how we *understand and view* the graph, it is a simple framework within which to design algorithms.

Although design sampling algorithms with a static model of computation in mind is indeed appropriate for some applications, it assumes the input graphs are relatively small, can fit entirely

into main memory, and have static structure (*i.e.*, not changing over the time). This is unrealistic for many domains. For instance, many social, communication, and information networks naturally change over time and are massive in size (*e.g.*, Facebook, Twitter, Flickr). The sheer size and dynamic nature of these networks make it difficult to load the full graph entirely in the main memory. Therefore, the static model of computation cannot realistically capture all the intricacies of graphs as we understand them today.

Many real-world networks that are currently of interest are *too large* to fit into memory. In this case, sampling methods that require random disk accesses can incur large I/O costs for loading and reading the data. Naturally, this raises a question as to how we can sample from large networks sequentially rather than assuming random access (*e.g.*, representing the graph as a stream of edges that is accessed in sequence)? In this context, most of the topology based sampling procedures such as breadth-first search and random-walk sampling are no longer appropriate as they require the ability to randomly access a node's neighbors $\mathcal{N}(v_i)$. If access is restricted to sequential passes over the edges, a large number of passes over the edges would be needed to repeatedly select $\mathcal{N}(\cdot)$. In a similar way, node sampling would no longer be appropriate as it not only requires random access for querying a node's neighbors but it also requires random access to the entire node set V in order to obtain a uniform random sample.

A *streaming model of computation* in which the graph can only be accessed sequentially as a stream of edges, is therefore more preferable for these situations [Zhang 2010]. The streaming model completely discards the possibility of random access to G and the graph can only be accessed through an ordered scan of the edge stream. The sampling algorithm may use the main memory for holding a portion of the edges temporarily and perform random accesses on that subset. In addition, the sampling algorithm may access the edges repeatedly by making multiple passes over the graph stream. Formally, for any input network G , we assume G arrives as a graph stream (as shown in Figure 2).

Definition 3.1 (Graph Stream). A *graph stream* is an ordered sequence of edges $e_{\pi(1)}, e_{\pi(2)}, \dots, e_{\pi(M)}$, where π is any random permutation on the edge indices $[M] = \{1, 2, \dots, M\}$, $\pi : [M] \rightarrow [M]$.

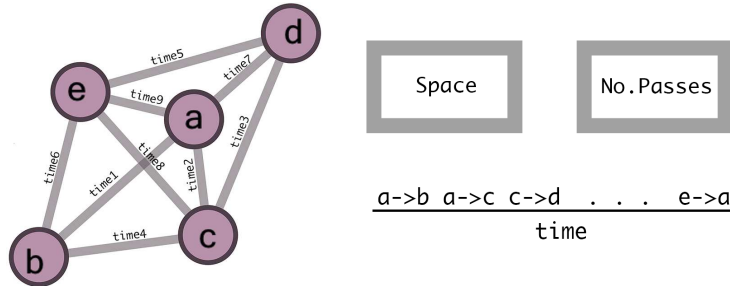


Fig. 2: Illustration of a *graph stream*---a sequence of edges ordered over time.

Definition 3.1 is usually called the "adjacency stream" model in which the graph is presented as a stream of edges in any arbitrary order. In contrast, the "incidence stream" model assumes all edges incident to a vertex are presented in order successively [Buriol et al. 2006]. In this paper, we assume the adjacency stream model.

While most real-world networks are too large to fit into main memory, many are also likely to be naturally *streaming*. A streaming graph is a continuous, unbounded, time-varying, transient stream of edges that is both too large and too dynamic to fit into memory. These types of streaming

graphs occur frequently in real-world communication and information domains. For example real-time tweets between users in Twitter, email logs, IP traffic, sensor networks, web search traffic, and many other applications. While sampling from these streaming networks is clearly challenging, their characteristics preclude the use of static models of computation thus a more systematic investigation is warranted. This naturally raises the second question: how can we sample from large graph streams in a *single pass* over the edges? Generally graph streams differ from static graphs in three main aspects:

- (1) The massive volume of edges streaming over the time is far too large to fit in main memory.
- (2) The graph can only be accessed sequentially in a single pass (*i.e.*, no random access to neighboring nodes or to the entire graph).
- (3) Efficient, real-time processing is of critical importance.

In a streaming model, as each edge $e \in E$ arrives, the sampling algorithm σ needs to decide whether to include the edge or not as the edge *streams* by. The sampling algorithm σ may also maintain state Ψ and consult the state to determine whether to sample e or not. The complexity of a streaming sampling algorithm is measured by:

- (1) Number of passes over the stream ω .
- (2) Space required to store the state Ψ and the output.
- (3) Representativeness of the output sample S .

Multiple passes over the stream (*i.e.*, $\omega > 1$) may be allowed for massive disk-resident graphs but multiple passes are not realistic for datasets where the graph is continuously streaming over time. In this case, a requirement of a single pass is more suitable (*i.e.*, $\omega = 1$). The total storage space (*i.e.*, Ψ) is usually of the order of the size of the output (*i.e.*, G_s): $|\Psi| = O(|G_s|)$. Note that this requirement is potentially larger than the $o(N, t)$ (and preferably $\text{polylog}(N, t)$) that streaming algorithms typically require [Muthukrishnan 2005]. But, since any algorithm cannot require less space than its output, we relax this requirement in our definition as follows.

Definition 3.2 (Streaming Graph Sampling). A *streaming graph sampling algorithm* is any sampling algorithm σ that produces a sampled graph G_s by sampling edges of the input graph G in a sequential order, preferably in *one pass* (*i.e.*, $\omega = 1$), while preferably maintaining state Ψ such that $|\Psi| \leq O(|G_s|)$.

Clearly, it is more difficult to design sampling algorithms for the *streaming graph model*, but it is critical to address the fundamental intricacies of graphs as we understand them today.

We now have what can be viewed as a complete spectrum of computational models for network sampling, which ranges from the simple, yet least realistic, static graph model to the more complex, but more realistic, streaming model (as in Figure 1). In the next sections, we evaluate algorithms for representative subgraph sampling in each computation model across the spectrum.

We note that our assumption in this work is that the population graph G is visible in its entirety (*collected and stored on disk*). In many analysis domains this assumption is valid, but in some cases the full structure of the population graph may be unknown prior to the sampling process (*e.g.*, the deep Web or distributed information in peer-to-peer networks). Web/network crawling is used extensively to sample from graphs that are not fully visible to the public but naturally allow methods to explore the neighbors of a given node (*e.g.*, hyperlinks in a web page). Topology-based sampling methods (*e.g.*, breadth-first search, random walk) have been widely used in this context. However, many of these methods assume the graph G is *well connected* and remains *static* during crawling, as discussed in [Gjoka et al. 2011].

4. RELATED WORK

Generally speaking, there are two bodies of work related to this paper: (i) network sampling methods, investigating and evaluating sampling methods with different goals of collecting a sample and

(ii) graph streams, mining graph streams. In this section, we describe and put the related work in perspective of the framework we discussed in Section 2.

The problem of sampling graphs has been of interest in many different fields of research. Most of this body of research has focused on *how* to sample and evaluate the goodness of the sample relative to the specific goal of the research.

Network sampling in social science. In social science, the classic work done by Frank in [Frank 1977] and his review papers in [Frank 1980] and [Frank 1981] provide the basic solutions to the first problems that arise when only a sample of the actors in a social network is available. Also, in [Goodman 1961], the concept of “chain-referral” sampling originated when Goodman introduced the snowball sampling method. Further, Granovetter introduced the network community to the problem of making inferences about the entire population from a sample (*e.g.*, estimation of network density) [Granovetter 1976]. And then later, respondent-driven sampling was proposed in [Heckathorn 1997] and analyzed in [Gile and Handcock 2010] to reduce the biases associated with chain referral sampling of hidden populations. For an excellent survey about estimation of sample properties, we refer the reader to the work in [Kolaczyk 2009]. Generally, the work in this area focuses on either the estimation of global network parameters (*e.g.*, density) or the estimation of actors (node) attributes, *i.e.*, goals 1 and 3.

Statistical properties of network sampling. Another important trend of research focused on analyzing the statistical properties of sampled subgraphs. For example, the work in [Lee et al. 2006; Yoon et al. 2007] studied the statistical properties of sampled subgraphs produced by the classical node, edge and random walk sampling and discussed the bias in estimates of topological properties. Similarly, the work done in [Stumpf et al. 2005] show that the sampled subgraph of a scale free network is far from being scale free. Conversely, the work done in [Lakhina et al. 2003] shows that under traceroute sampling, the degree distribution is a Power law even when the actual distribution is a Poisson. Clearly, the work in this area has focused on representative subgraph sampling (*i.e.*, goal 2) considering how sampling changes the topological properties of the original network.

Network sampling in network systems research. A large body of research in network systems focused on Internet measurement, which targets the problem of topology measurements in large-scale online networks, such as peer-to-peer networks (P2P), world wide web (WWW), and online social networks (OSN). The sheer size, and distributed structure of these networks make it hard to measure the properties of the entire network. Network sampling, via *crawling*, has been used extensively in this context. In OSNs, sampling methods that don’t allow nodes to be revisited are widely used (*e.g.*, breadth-first search [Ahn et al. 2007; Mislove et al. 2007; Wilson et al. 2009]). Breadth-first search, however, has been shown to be biased towards high degree nodes [Ye et al. 2010] but the work in [Kurant et al. 2011] suggested analytical solutions to correct the bias. Random walk sampling has also been used, such as the work in [Gjoka et al. 2010] to sample a uniform sample from users in Facebook, and Last.fm. For a recent survey covering assumptions and comparing different methods of crawling, we refer the reader to [Gjoka et al. 2011]. Similar to OSNs, random walk sampling and its variants were used extensively to sample the WWW [Baykan et al. 2009; Henzinger et al. 2000], and P2P networks [Gkantsidis et al. 2004]. Since the classical random walk is biased towards high degree nodes, some improvements were applied to correct the bias. For example, the work in [Stutzbach et al. 2006] applied metropolis-hastings random walk (MHRW) to sample peers in Gnutella network, and the work in [Rasti et al. 2009] applied re-weighted random walk (RWRW) to sample P2P networks. Other work used *m*-dependent random walk and random walks with jumps [Ribeiro and Towsley 2010; Avrachenkov et al. 2010].

Overall, the work done in this area has focused extensively on sampling a uniform subset of nodes from the graph, to estimate topological properties of the entire network from the set of sampled nodes (*i.e.*, goal 1).

Network sampling in structured data mining. Network sampling is a core part of data mining research. Representative subgraph sampling was first defined in [Leskovec and Faloutsos 2006]. Instead of sampling, the work in [Krishnamurthy et al. 2007] explored reductive methods to shrink the existing topology of the graph. Further, the work in [Hubler et al. 2008] proposed a generic metropolis algorithm to optimize the representativeness of a sampled subgraph (by minimizing the distance of several graph properties). Unfortunately, the number of steps until convergence is not known in advance (and usually large), and each step requires the computation of a distance function, which may be costly. However, other work discussed the difficulty of getting a "universal representative" subgraph that preserves *all* properties of the target network. For example, our work in [Ahmed et al. 2010b] discussed the possible correlation between properties, where accurately preserving some properties leads to underestimate/overestimate other properties (*e.g.*, preserving average degree of the target network leads to overestimating its density). Also, the work done in [Maiya and Berger-Wolf 2011] investigated the connection between the biases of topology-based sampling methods (*e.g.*, breadth-first search) and some topological properties (*e.g.*, degree). Therefore, some work focused on obtaining samples for specific applications, and to satisfy specific properties of the target network, such as to preserve the community structure [Maiya and Berger-Wolf 2010], to preserve the pagerank between all pairs of sampled nodes [Vattani et al. 2011], and to visualize the graph [Jia et al. 2008].

Other network sampling goals have been considered as well. For example, sampling nodes to perform A/B testing of social features [Backstrom and Kleinberg 2011], sampling nodes to analyze estimators of the fraction of users with a certain property [Dasgupta et al. 2012], (*i.e.*, goal 3), and sampling tweets (edges) to analyze the language used in twitter (*i.e.*, goal 4). In addition, [Al Hasan and Zaki 2009] samples the output space of graph mining algorithms, [Papagelis et al. 2011] collects information from social peers for enhancing the information needs of users, and [De Choudhury et al. 2010] studies the impact of sampling on the discovery of information diffusion.

Much of this work has focused on sampling in the static model of computation, where it is assumed that the graph can be loaded entirely in main memory, or the graph is distributed and allows exploring the neighborhood of nodes in a crawling fashion.

Graph Streams. Data stream querying and mining has gained a lot of interest in the past years [Babcock et al. 2002a; Golab and Özsu 2003; Muthukrishnan 2005; Aggarwal 2006a]. For example, for sequence sampling (*e.g.*, reservoir sampling) [Vitter 1985; Babcock et al. 2002b; Aggarwal 2006b], for computing frequency counts [Manku and Motwani 2002; Charikar et al. 2002] and load shedding [Tatbul et al. 2003], for mining concept drifting data streams [Wang et al. 2003; Gao et al. 2007; Fan 2004b; Fan 2004a], clustering evolving data streams [Guha et al. 2003; Aggarwal et al. 2003], active mining and learning in data streams [Fan et al. 2004; Li et al. 2009], and other related mining tasks [Domingos and Hulten 2000; Hulten et al. 2001; Gaber et al. 2005; Wang et al. 2005].

Recently, there has been an increasing interest in mining and querying *graph streams* as a result of the proliferation of graph data (*e.g.*, social networks, emails, IP traffic, Twitter hashtags). Following the earliest work on graph streams [Henzinger et al. 1999], various problems were explored in the field of mining graph streams. For example, to count triangles [Bar-Yossef et al. 2002; Buriol et al. 2006], finding common neighborhoods [Buchsbaum et al. 2003], estimating pagerank values [Sarma et al. 2008], and characterizing degree sequences in multi-graph streams [Cormode and Muthukrishnan 2005]. More recently, there is the work done on clustering graph streams [Aggarwal et al. 2010b], outlier detection [Aggarwal et al. 2011], searching for subgraph patterns [Chen and Wang 2010], and mining dense structural patterns [Aggarwal et al. 2010a].

Graph stream sampling was utilized in some of the work mentioned above. For example, the work in [Sarma et al. 2008] performs short random walks from uniformly sampled nodes to estimate pagerank scores. Also, [Buriol et al. 2006] used sampling to estimate number of triangles in the

graph stream. Moreover, the work in [Cormode and Muthukrishnan 2005] used a min-wise hash function to sample nearly uniformly from the set of all edges that has been at any time in the stream. The sampled edges were later used to maintain cascaded summaries of the graph stream. More recently, [Aggarwal et al. 2011] designed a structural reservoir sampling approach (based on min-wise hash sampling of edges) for structural summarization. For an excellent survey on mining graph streams, we refer the reader to [McGregor 2009; Zhang 2010].

The majority of this work has focused on sampling a subset of nodes uniformly from the stream to estimate parameters such as the number of triangles or pagerank scores of the graph stream (*i.e.* goal 1). Also as we discussed above, other work has focused on sampling a subset of edges uniformly from the graph stream to maintain summaries (*i.e.* goal 2). These summaries can be further pruned (in the decreasing order of the hash value [Aggarwal et al. 2011]) to satisfy a specific stopping constraint (*e.g.* specific number of nodes in the summary). In this paper, since we focus primarily on sampling a representative subgraph $G_s \subset G$ from the graph stream, we compare to some of these methods in Section 6.

5. SAMPLING FROM STATIC GRAPHS

In this section, we focus on how to sample a representative subgraph $G_s = (V_s, E_s)$ from $G = (V, E)$ (*i.e.*, goal 2 from Section 2.2). A representative sample G_s is essential for many applications in machine learning, data mining, and network simulations. As an example, it can be used to drive realistic simulations and experimentation before deploying new protocols and systems in the field [Krishnamurthy et al. 2007]. We evaluate the representativeness of G_s relative to G , by comparing distributions of six topological properties calculated over nodes, edges, and subgraphs (as summarized in Table I).

First, we distinguish between the degree of the sampled nodes before and after sampling. For any node $v_i \in V_s$, we denote k_i to be the node degree of v_i in the input graph G . Similarly, we denote k_i^s to be the node degree of v_i in the sampled subgraph G_s . Note that $k_i = |\mathcal{N}(v_i)|$, where $\mathcal{N}(v_i) = \{v_j \in V | e_{ij} = (v_i, v_j) \in E\}$ is the set of neighbors of node v_i . Clearly, when a node is sampled, it is not necessarily the case that all its neighbors are sampled as well, and therefore $0 \leq k_i^s \leq k_i$.

In this section, we propose a simple and efficient sampling algorithm based on the concept of graph-induction: *induced edge sampling* (for brevity ES-i). ES-i has several advantages over current sampling methods as we show later in this section:

- (1) ES-i preserves the topological properties of G better than many of current sampling algorithms.
- (2) ES-i can be easily implemented as a *streaming* sampling algorithm using only two passes over the edges of G (*i.e.*, $\omega = 2$).
- (3) ES-i is suitable for sampling large graphs that cannot fit into main memory.

We compare our proposed algorithm ES-i to state-of-the-art sampling algorithms from each of the three classes of network sampling (node, edge, and topology-based). More specifically, we compare to node (NS), edge (ES), and forest fire (FFS) sampling methods. Note that all the baseline methods are implemented under the assumption of a static model of computation. However, we show how ES-i can be implemented as a streaming algorithm that takes only two passes over the edges of G (*i.e.* $\omega = 2$). Thus its computational complexity is $O(2E)$.

5.1. Algorithm

We formally specify ES-i in Algorithm 1. Initially, ES-i selects the nodes in pairs by sampling edges uniformly (*i.e.*, $p(e_{ij} \text{ is selected}) = 1/|E|$) and adds them to the sample (V_s). Then, ES-i augments the sample with all edges that exist between any of the sampled nodes ($E_s = \{e_{ij} = (v_i, v_j) \in E | v_i, v_j \in V_s\}$). These two steps together form the sample subgraph $G_s = (V_s, E_s)$. For example, suppose edges $e_{12} = (v_1, v_2)$ and $e_{34} = (v_3, v_4)$ are sampled in the first step, that leads to the addition of the vertices v_1, \dots, v_4 into the sampled graph. In the second step, ES-i adds all the edges that exist between the sampled nodes---for example, edges $e_{12} = (v_1, v_2)$, $e_{34} = (v_3, v_4)$,

$e_{13} = (v_1, v_3)$, $e_{24} = (v_2, v_4)$, and any other possible combinations involving v_1, \dots, v_4 that appear in G .

ALGORITHM 1: ES-i(ϕ, E)

Input : Sample fraction ϕ , Edge set E
Output : Sampled Subgraph $G_s = (V_s, E_s)$

```

1  $V_s = \emptyset, E_s = \emptyset$ 
2 // Node selection step
3 while  $|V_s| < \phi \times |V|$  do
4    $r = \text{random}(1, |E|)$ 
5   // uniformly random
6    $e_r = (u, v)$ 
7    $V_s = V_s \cup \{u, v\}$ 
8 // Edge selection step
9 for  $k = 1 : |E|$  do
10   $e_k = (u, v)$ 
11  if  $u \in V_s$  AND  $v \in V_s$  then
12     $E_s = E_s \cup \{e_k\}$ 

```

Downward bias caused by sampling. Since any sampling algorithm (by definition) selects only a subset of the nodes/edges in the graph G , it naturally produces subgraphs with underestimated degrees in the degree distribution of G_s . We refer to this as a *downward bias* and note that it is a property of all network sampling methods, since only a fraction of a node's neighbors may be selected for inclusion in the sample (*i.e.* $k_i^s \leq k_i$ for any sampled node v_i).

Our proposed sampling methods exploits two key observations. First, by selecting nodes via edge sampling the method is inherently biased towards the selection of nodes with high degrees, resulting in an *upward bias* in the (original) degree distribution if only observed from the sampled nodes (*i.e.*, using the degree of the sampled nodes as observed in G). The upward bias resulting from edge sampling can help offset the downward bias of the sampled degree distribution of G_s .

In addition to improving estimates of the sampled degree distribution, selecting high degree nodes also helps to produce a more connected sample subgraph that preserves the topological properties of the graph G . This is due to the fact that high degree nodes often represent *hubs* in the graph, which serve as good navigators through the graph (*e.g.*, many shortest paths usually pass through these hub nodes).

However, while the upward bias of edge sampling can help offset some issues of sample selection bias, it is not sufficient to use it in isolation to construct a good sampled subgraph. Specifically, since the edges are each sampled independently, edge sampling is unlikely to preserve much structure *surrounding* each of the selected nodes. This leads us to our second observation, that a simple *graph induction* step over the sampled nodes (where we sample all the edges between any sampled nodes from G) is a key to recover much of the connectivity in the sampled subgraph---offsetting the downward degree bias as well as increasing local clustering in the sampled graph. More specifically, graph induction increases the likelihood that triangles will be sampled among the set of selected nodes, resulting in higher clustering coefficients and shorter path lengths in G_s .

These observations, while simple, make the sample subgraph G_s approximate the characteristics of the original graph G more accurately, even better than topology-based sampling methods.

Sampling very large graphs. As we discussed before, many real networks are now too large to fit into main memory. This raises the question: how can we sample from G sequentially, one edge at a time, while minimizing the number of passes over the edges? As we discussed in Section 3, most of the topology-based sampling methods would no longer be applicable, since they require

many passes over the edges. Conversely, ES-i runs sequentially and requires only two passes over the edges of G (i.e., $\omega = 2$). In the first pass, ES-i samples edges uniformly from the stream of edges E and adding their incident nodes to the sample. Then, it proceeds in a second pass by adding all edges which have both end-points already in the sample. This makes ES-i suitable for sampling large graphs that cannot fit into main memory.

Next, we analyze the characteristics of ES-i and after that, in the evaluation, we show how it accurately preserves many of the properties of the graph G .

5.2. Analysis of ES-i

In this section, we analyze the bias of ES-i's node selection analytically by comparing to the unbiased case of uniform sampling in which all nodes are sampled with a uniform probability (i.e., $p = \frac{1}{N}$). First, we denote f_D to be the degree sequence of G where $f_D(k)$ is the number of nodes with degree k in graph G . Let $n = |V_s|$ be the target number of nodes in the sample subgraph G_s (i.e., $\phi = \frac{n}{N}$).

Upward bias to high degree nodes. We start by analyzing the upward bias to select high degree nodes by calculating the expected value of the number of nodes with original degree k that are added to the sample set of nodes V_s . Let $E[f_D(k)]$ be the expected value of $f_D(k)$ for the sampled set V_s when n nodes are sampled uniformly with probability $p = \frac{1}{N}$:

$$\begin{aligned} E[f_D(k)] &= f_D(k) \cdot n \cdot p \\ &= f_D(k) \cdot \frac{n}{N} \end{aligned}$$

Since, ES-i selects nodes proportional to their degree, the probability of sampling a node v_i with degree $k_i = k$ is $p' = \frac{k}{\sum_{j=1}^N k_j}$. Note that we can also express the probability as $p' = \frac{k}{2 \cdot |E|}$. Then we let $E'[f_D(k)]$ denote the expected value of $f_D(k)$ for the sampled set V_s when nodes are sampled with ES-i:

$$\begin{aligned} E'[f_D(k)] &= f_D(k) \cdot n \cdot p' \\ &= f_D(k) \cdot n \cdot \frac{k}{2 \cdot |E|} \end{aligned}$$

This leads us to Lemma 5.1.

LEMMA 5.1. *ES-i is biased to high degree nodes.*

$E'[f_D(k)] \geq E[f_D(k)]$ if $k \geq k_{avg}$, where $k_{avg} = \frac{2 \cdot |E|}{N}$ is the average degree in G .

PROOF: Consider the threshold k at which the expected value of $f_D(k)$ using ES-i sampling is greater than the expected value of $f_D(k)$ using uniform random sampling:

$$\begin{aligned} E'[f_D(k)] - E[f_D(k)] &\geq 0 \\ f_D(k) \cdot n \cdot \frac{k}{2 \cdot |E|} - f_D(k) \cdot \frac{n}{N} &\geq 0 \\ f_D(k) \cdot \frac{n}{N} \cdot \frac{k}{k_{avg}} - f_D(k) \cdot \frac{n}{N} &\geq 0 \\ \frac{k}{k_{avg}} - 1 &\geq 0 \\ \frac{k}{k_{avg}} &\geq 1 \end{aligned}$$

The above statement is true when $k \geq k_{avg}$. \square

Downward bias caused by sampling. Next, instead of focusing on the *original* degree k as observed in the graph G , we focus on the *sampled* degree k^s as observed in the sample subgraph G_s , where $0 \leq k^s \leq k$. Let k_i^s be a random variable that represent the sampled degree of node v_i in G_s , given that the original degree of node v_i in G was k_i . We compare the expected value of k_i^s when using uniform sampling to the expected value of k_i^s when using ES-i. Generally, the degree of the node v_i in G_s depends on how many of its neighbors in G are sampled. When using uniform sampling, the probability of sampling one of the node's neighbors is $p = \frac{1}{N}$:

$$E[k_i^s] = \sum_{j=1}^{k_i} p \cdot n = k_i \cdot \frac{n}{N}$$

When using ES-i, the probability of sampling any of the node's neighbors is proportional to the degree of the neighbor. Let v_j be a neighbor of v_i (i.e., $e_{ij} = (v_i, v_j) \in E$), then the probability of sampling v_j is $p' = \frac{k_j}{2 \cdot |E|}$:

$$\begin{aligned} E'[k_i^s] &= \sum_{j=1}^{k_i} p' \cdot n \\ &= \sum_{j=1}^{k_i} \frac{k_j}{2 \cdot |E|} \cdot n \\ &= \frac{n}{N} \cdot \frac{\sum_{j=1}^{k_i} k_j}{k_{avg}} \end{aligned}$$

Now, let us define the variable $k_{\mathcal{N}} = \sum_{k'} k' \cdot P(k'|k)$, where $k_{\mathcal{N}}$ represents the average degree of the neighbors of a node with degree k as observed in G . The function $k_{\mathcal{N}}$ has been widely used as a global measure of the *assortativity* in a network [Newman 2002]. If $k_{\mathcal{N}}$ is increasing with k , then the network is assortative---indicating that nodes with high degree connect to, on average, other nodes with high degree. Alternatively, if $k_{\mathcal{N}}$ is decreasing with k , then the network is dissortative--indicating that nodes of high degree tend to connect to nodes of low degree.

Note that here, we define $k_{\mathcal{N}i} = \frac{\sum_{j=1}^{k_i} k_j}{k_i}$ as the average degree of the neighbors of node v_i . Note that $k_{\mathcal{N}i} \geq 1$. In this context, $k_{\mathcal{N}i}$ represents the local assortativity of node v_i , so then:

$$\begin{aligned} E'[k_i^s] &= \frac{n}{N} \cdot \frac{\sum_{j=1}^{k_i} k_j}{k_{avg}} \\ &= \frac{n}{N} \cdot \frac{k_i}{k_{avg}} \cdot \frac{\sum_{j=1}^{k_i} k_j}{k_i} \\ &= k_i \cdot \frac{n}{N} \cdot \frac{k_{\mathcal{N}i}}{k_{avg}} \end{aligned}$$

This leads us to Lemma 5.2.

LEMMA 5.2. *The expected sampled degree in V_s using ES-i is greater than the expected sampled degree based on uniform node sampling. For any node $v_i \in V_s$, $E'[k_i^s] \geq E[k_i^s]$ if $k_{\mathcal{N}i} \geq k_{avg}$,*

where the average degree in G is $k_{avg} = \frac{2 \cdot |E|}{N}$ and the average degree of v_i 's neighbors in G is $k_{\mathcal{N}_i} = \frac{\sum_{j=1}^{k_i} k_j}{k_i}$.

PROOF: Consider the threshold k at which the expected value of k^s using ES-i sampling is greater than the expected value of k^s using uniform random sampling:

$$\begin{aligned} E'[k_i^s] - E[k_i^s] &\geq 0 \\ k_i \cdot \frac{n}{N} \cdot \frac{k_{\mathcal{N}_i}}{k_{avg}} - k_i \cdot \frac{n}{N} &\geq 0 \\ \frac{k_{\mathcal{N}_i}}{k_{avg}} - 1 &\geq 0 \\ \frac{k_{\mathcal{N}_i}}{k_{avg}} &\geq 1 \end{aligned}$$

The above statement is true when $k_{\mathcal{N}_i} \geq k_{avg}$. \square

Generally, for any sampled node v_i with degree k_i , if the average degree of v_i 's neighbors is greater than the average sampled degree of G , then the expected sampled degree k_i^s under ES-i is higher than if uniform sampling is used. This typically the case in many real networks where high degree nodes are connected with other high degree nodes.

In Figure 3, we empirically investigate the difference between the sampled degrees k^s and original degrees k for an example network---the CONDMAT graph. Specifically we compare the degree distribution of G_s (in Figure 3b) to the degree distribution of G (in Figure 3a) as observed only from the set of sampled nodes V_s . Furthermore, we compare to the *full* degree distribution of G , *i.e.*, as observed over the set of all nodes V .

In Figure 3a, NS accurately estimates the actual degree distribution as observed in G . However, in Figure 3b, NS underestimates the sampled degree distribution in G_s , *i.e.*, the NS curve is skewed upwards. On the other hand, in Figure 3a, ES, FFS, and ES-i overestimate the original degree distribution in G , since they are biased to selecting higher degree nodes. In Figure 3b, ES and FFS both underestimate the sampled degree distribution as observed in G_s . In contrast to other sampling methods, ES-i comes closer to replicating the original degree distribution of G in G_s . We conjecture that the combination of high degree bias together with the graph induction helps ES-i to compensate for the underestimation caused by sampling subgraphs.

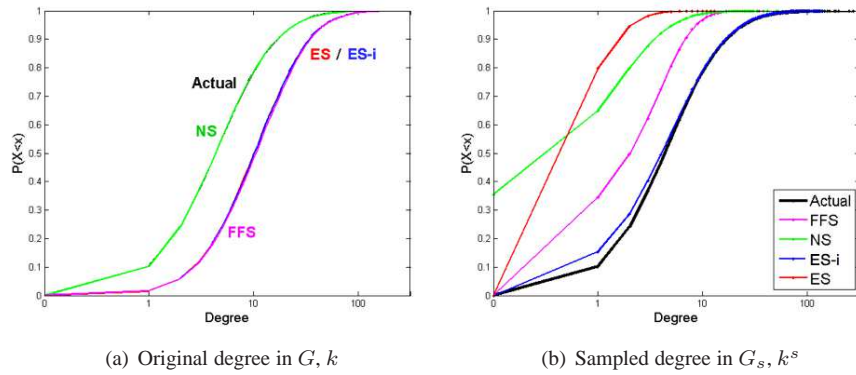


Fig. 3: Illustration of original degrees (in G) vs. sampled degrees (in G_s) for subgraphs selected by NS, ES, ES-i, and FFS on the CondMAT network.

Table II: Characteristics of Network Datasets

Graph	Nodes	Edges	Weak Comps.	Avg. Path	Density	Global Clust.
HEPPH	34,546	420,877	61	4.33	7×10^{-4}	0.146
CONDMAT	23,133	93,439	567	5.35	4×10^{-4}	0.264
TWITTER	8,581	27,889	162	4.17	7×10^{-4}	0.061
FACEBOOK	46,952	183,412	842	5.6	2×10^{-4}	0.085
FLICKR	820,878	6,625,280	1	6.5	1.9×10^{-5}	0.116
LIVEJOURNAL	4,847,571	68,993,773	1876	6.5	5.8×10^{-6}	0.2882
EMAIL-PU UNIV	214,893	1,270,285	24	3.91	5.5×10^{-5}	0.0018

5.3. Experimental Evaluation

In this section, we present results evaluating the various sampling methods on static graphs. We compare the performance of our proposed algorithm ES-i to other algorithms from each class (as discussed in section 2.3): node sampling (NS), edge sampling (ES) and forest fire sampling (FFS). We compare the algorithms on seven different real-world networks. We use online social networks from FACEBOOK from New Orleans City [Viswanath et al. 2009] and FLICKR [Gleich 2012]. Social media networks drawn from TWITTER, corresponding to users tweets surrounding the United Nations Climate Change Conference in Copenhagen, December 2009 (*#cop15*) [Ahmed et al. 2010a]. Also, we use a citation graph from ArXiv HEPH, and a collaboration graph from CONDMAT [SNAP]. Additionally, we use an email network EMAIL-UNIV that corresponds to a month of email communication collected from Purdue university mail-servers [Ahmed et al. 2012]. Finally, we compare the methods on a large social network from LIVEJOURNAL [SNAP] with 4 million nodes (included only at the 20% sample size). Table II provides a summary of the global statistics of these six network datasets.

Next, we outline the experimental methodology and results. For each experiment, we apply to the full network and sample subgraphs over a range of sampling fractions $\phi = [5\%, 40\%]$. For each sampling fraction, we report the average results over ten different trials.

Distance metrics. Figures 4(a)–4(d) show the average KS statistic for degree, path length, clustering coefficient, and k-core distributions on the six datasets. Generally, ES-i outperforms the other methods for each of the four distributions. FFS performs similar to ES-i in the degree distribution, however, it does not perform well for path length, clustering coefficient, and k-core distributions. This implies that FFS can capture the degree distribution but not connectivity between the sampled nodes. NS performs better than FFS and ES for path length, clustering coefficient, and k-core statistics but not for the degree statistics. This is due to the uniform sampling of the nodes that makes NS is more likely to sample low degree nodes and miss the high degree nodes (as discussed in 5.2). Clearly, as the sample size increases, NS is able to select more nodes and thus the KS statistic decreases. ES-i and NS perform similarly for path length distribution. This is because they both form a fully induced subgraph out of the sampled nodes. Since induced subgraphs are more connected, the distance between pairs of nodes is generally smaller in induced subgraphs.

In addition, we also used skew divergence as a second measure. Figures 4(e)–4(h) show the average skew divergence statistic for degree, path length, clustering coefficient, and k-core distributions on the six datasets. Note that skew divergence computes the divergence between the sampled and the real distributions on the entire support of the distributions. While the skew divergence shows that ES-i outperforms the other methods similar to KS statistic, it also shows that ES-i performs significantly better across the entire support of the distributions.

Finally, Figures 4(i) and 4(j) show the L1 and L2 distances for eigenvalues and network values respectively. Clearly, ES-i outperforms all the other methods that fail to improve their performance even when the sample size is increased up to 40% of the full graph.

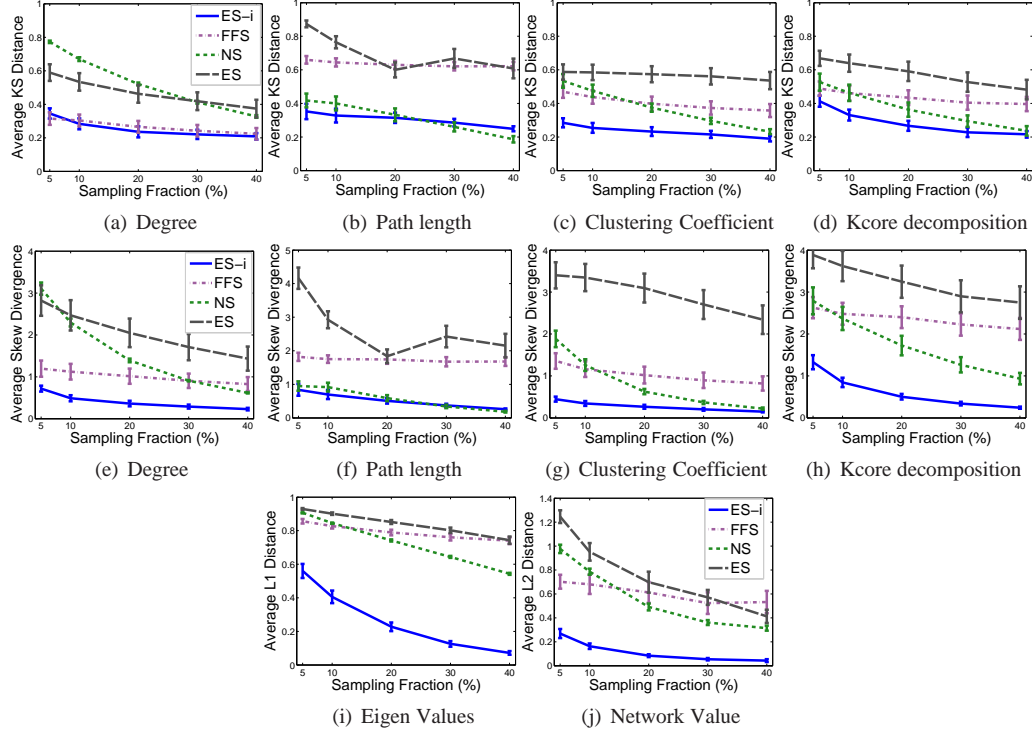


Fig. 4: (a-d) Average KS distance, (e-h) average skew divergence, and (i-j) average L1 and L2 distance respectively, across 6 datasets.

Distributions. While the distance measures are important to quantify the divergence between the sampled and the real distributions, by analyzing only the distance measures it is unclear whether the sampled statistics are an over-estimate or under-estimate of the original statistics. Therefore, we plot the distributions of for all networks at the 20% sample size. We choose the 20% as a representative sample size, however, we note that same conclusions hold for the other sample sizes. Note that we plot the *CCDF* for degree and k-core distributions, *CDF* for path length and clustering coefficient distribution, and we plot eigenvalues and network values versus the rank. Figures 10, 12, 11, 13, 14, 15, and 16 (Appendix A) show the plots for all the distributions across all networks.

Degree Distribution. Across all networks, ES-i captures the tail of the degree distribution (high degree nodes) better than NS, ES, and FFS. However, ES-i under-estimates the low degree nodes for TWITTER, EMAIL-UNIV, and FLICKR. FFS and NS capture a large fraction of low degree nodes but they fail to capture the high degree nodes.

Path length Distribution. ES-i preserves the path length distribution of HEPH, CONDMAT, and LIVEJOURNAL, however, it underestimates the distributions of TWITTER, EMAIL-UNIV, and FLICKR. Conversely, NS over-estimates the distributions of HEPH, CONDMAT, and LIVEJOURNAL but successfully preserves the distributions of the other datasets.

Clustering Coefficient Distribution. ES-i generally captures the clustering coefficient more accurately than other methods. While ES-i under-estimates the low clustering coefficients specially in EMAIL-UNIV, and FLICKR, the other methods fail to capture the clustered structures in almost all the datasets.

K-Core Distribution. Similar to the previous statistics, ES-i nicely preserves the distribution of core sizes for HEPH, CONDMAT, and FACEBOOK, but it over-estimates the core structures of the other datasets. On the other hand, NS, ES, and FFS generally fail to capture the core structures for the majority of the datasets (except FLICKR). In addition to the distribution of the core sizes, we compared the *max-core* number in the sampled graphs to their real counterparts for the 20% sample size (Table III). Note that the *max-core number* is the maximum value of k in the k -cor distribution. In contrast to ES-i, the *max-core number* of NS, ES, and FFS is consistently an order of magnitude smaller than the real *max-core number*. This indicates that NS, ES, and FFS do not preserve the local density in the sampled subgraph structures.

Eigenvalues. The NS, ES, and FFS methods generally fail to approximate the eigenvalues of the original graph in the sample. Conversely, ES-i accurately approximates the eigenvalues of TWITTER, EMAIL-UNIV, FLICKR, and LIVEJOURNAL and closely approximates the eigenvalues of HEPH, CONDMAT, and FACEBOOK (at 20% sample size). By the *interlacing* theorem of the eigenvalues of induced subgraphs [Haemers 1995], the eigenvalues of ES-i in G_s can be used to estimate bounds on their counterparts in the input graph G : $\lambda_i \leq \mu_i \leq \lambda_{i+(N-n)}$ such that μ_i is the i^{th} eigenvalue of G_s , and λ_i is the i^{th} eigenvalue of G .

Network values. Similar to the other graph measures, ES-i accurately approximates the network values of the graph compared to other methods.

Table III: Comparison of *max-core-number* at the 20% sample size for ES-i, NS, ES, FFS versus Real value of G .

Graph	Real max core no.	ES-i	NS	ES	FFS
HEPPH	30	23*	8	2	4
CONDMAT	25	20*	7	2	6
TWITTER	18	18*	5	2	3
FACEBOOK	16	14*	4	2	3
FLICKR	406	406*	83	21	7
LIVEJOURNAL	372	372*	84	6	7
EMAIL-UNIV	47	46*	15	3	7

Summary---We summarize the main empirical conclusions in this section:

- (1) Sampled subgraphs collected and constructed by ES-i accurately preserve a range of network statistics that capture both local and global distributions of the graph structure.
- (2) Due to its bias to selecting high degree nodes, ES-i generally favors dense and clustered areas of the graph, which results in connected sample subgraphs---in contrast with other methods.
- (3) NS, ES and FFS generally construct more sparsely connected sample subgraphs.

6. SAMPLING FROM STREAMING GRAPHS

In this section, we focus on how to sample a representative subgraph G_s from G , such that G is presented as a stream of edges in no particular order. Note that in this paper we focus on space-efficient sampling methods. Using the definition of a streaming graph sampling algorithm as discussed in Section 3, we now present streaming variants of different sampling algorithms as discussed in Section 2.3.

Streaming Node Sampling. One key problem with traditional node sampling we discussed in 2.3 is that the algorithms assume that nodes can be accessed at random. In our stream setting, new nodes arrive into the system only when an edge that contains the new node is added into the system; it is therefore difficult to identify which n nodes to select *a priori*. To address this, we utilize the

idea of reservoir sampling [Vitter 1985] to implement a streaming variant of node sampling (see Algorithm 2).

The main idea is to select nodes uniformly at random with the help of a uniform random hash function. A uniform random hash function defines a true random permutation on the nodes in the graph, meaning that any node is equally likely to be the minimum. Specifically, we keep track of nodes with n smallest hash values in the graph; nodes are only added if their hash values represent the top- n minimum hashes among all nodes seen thus far in the stream. Any edge that has both vertices already in the reservoir is automatically added to the original graph.

Since the reservoir is finite, a node with smaller hash value may arrive late in the stream and replace a node that was sampled earlier. In this case, all edges incident to the node that was dropped will be removed from the sampled subgraph. Once the reservoir is filled up to n nodes, it will remain at n nodes, but since the selection is based on the hash value, nodes will be dropped and added as the algorithm samples from all portions of the stream (not just the front). Therefore, it guarantees a uniformly sampled set of nodes from the graph stream.

ALGORITHM 2: Streaming Node Sampling $NS(n, S)$

Input : Sample Size n , Graph Stream S
Output : Sampled Subgraph $G_s = (V_s, E_s)$

```

1  $V_s = \emptyset, E_s = \emptyset$ 
2  $h$  is fixed uniform random hash function
3  $t = 1$ 
4 for  $e_t$  in the graph stream  $S$  do
5    $(u, v) = e_t$ 
6   if  $u \notin V_s \& h(u)$  is top- $n$  min hash then
7      $V_s = V_s \cup u$ 
8     Remove all edges incident on replaced node
9   if  $v \notin V_s \& h(v)$  is top- $n$  min hash then
10     $V_s = V_s \cup v$ 
11    Remove all edges incident on replaced node
12   if  $u, v \in V_s$  then
13      $E_s = E_s \cup e$ 
14    $t = t + 1$ 

```

Streaming Edge Sampling. Streaming edge sampling can be implemented similar to streaming node sampling. Instead of hashing individual nodes, we focus on using hash-based selection of edges (as shown in Algorithm 3). We use the approach that was first proposed in [Aggarwal et al. 2011]. More precisely, if we are interested in obtaining m edges at random from the stream, we can simply keep a reservoir of m edges with the minimum hash value. Thus, if a new edge streams into the system, we check if its hash value is within top- m minimum hash values. If it is not, then we do not select that edge, otherwise we add it to the reservoir while replacing the edge with the previous highest top- m minimum hash value. One problem with this approach is that our goal is often in terms of sampling a certain number of nodes n . Since we use a reservoir of edges, finding the right m that provides n nodes is hard. It also keeps varying depending on which edges the algorithm ends up selecting. Note that sampling fraction could also be specified in terms of fraction of edges; the choice of defining it in terms of nodes is somewhat arbitrary in that sense. For our comparison purposes, we ensured that we choose a large enough m such that the number of nodes was much higher than n , but later iteratively pruned out sampled edges with the maximum hash values until the target number of nodes n was reached.

ALGORITHM 3: Streaming ES(n, S)

Input : Sample Size n , Graph Stream S
Output : Sampled Subgraph $G_s = (V_s, E_s)$

```

1  $V_s = \emptyset, E_s = \emptyset$ 
2  $h$  is fixed uniform random hash function
3  $t = 1$ 
4 for  $e_t$  in the graph stream  $S$  do
5    $(u, v) = e_t$ 
6   if  $h(e_t)$  is in top- $m$  min hash then
7      $E_s = E_s \cup e_t$ 
8      $V_s = V_s \cup \{u, v\}$ 
9   Iteratively remove edges in  $E_s$  in decreasing order such that  $|V_s| = n$  nodes
10   $t = t + 1$ 

```

Streaming Topology-Based Sampling. We also consider a streaming variant of a topology-based sampling algorithm. Specifically, we consider a simple BFS-based algorithm (shown in Algorithm 4) that works as follows. This algorithm essentially implements a simple breadth-first search on a sliding window of w edges in the stream. In many respects, this algorithm is similar to the forest-fire sampling (FFS) algorithm. Just as in FFS, it starts at a random node in the graph and selects an edge to burn (as in FFS parlance) among all edges incident on that node within the sliding window. For every edge burned, let v be the incident node at the other end of the burned edge. We enqueue v onto a queue Q in order to get a chance to burn its incident edges within the window. For every new streaming edge, the sliding window moves one step, which means the oldest edge in the window is dropped and a new edge is added. (If that oldest edge was sampled, it will still be part of the sampled graph.) If as a result of the sliding window moving one step, the node has no more edges left to burn, then the burning process will dequeue a new node from Q . If the queue is empty, the process jumps to a random node within the sliding window (just as in FFS). This way, it does BFS as much as possible within a sliding window, with random jumps if there is no more edges left to explore. Note that there may be other possible implementation of a streaming variant for topology-based sampling, but since to our knowledge, there are no streaming algorithms in the literature, we include this as a reasonable approximation for comparison. This algorithm has a similar problem as the edge sampling variant in that it is difficult to control the exact number of sampled nodes, and hence some additional pruning needs to be done at the end (see Algorithm 4).

Partially-Induced Edge Sampling (PIES). We finally present our main algorithm called PIES that outperforms the above implementation of stream sampling algorithms. Our approach discussed in section 5 outlines a sampling algorithm based on edge sampling concepts. A key advantage of using edge sampling is its bias towards high degree nodes. This upward bias helps offset the downward bias (caused by subgraph sampling) to some extent. Afterwards, forming the induced graph will help capture the connectivity among the sampled nodes.

Unfortunately, full graph induction in a streaming fashion is difficult, since node selection and graph induction requires at least two passes (when implemented in the obvious, straightforward way). Thus, instead of full induction of the edges between the sampled nodes, we can utilize *partial* induction and combine edge-based node sampling with the graph induction (as shown in Algorithm 5) in a single step. The partial induction step induces the sample in the forward direction only. In other words, it adds any edge among a pair of sampled nodes if it occurs *after* both the two nodes were added to the sample.

PIES aims to maintain a dynamic sample while the graph is streaming by utilizing the same reservoir sampling idea we have used before. In brief, we add the first m edges of the stream to a *reservoir* such that the reservoir contains n nodes and then the rest of the stream is processed randomly by replacing existing records in the reservoir. Specifically, PIES runs over the edges in

ALGORITHM 4: Streaming BFS($n, S, wsize$)**Input** : Sample Size n , Graph Stream S , Window Size= $wsize$ **Output** : Sampled Subgraph $G_s = (V_s, E_s)$

```

1  $V_s = \emptyset, E_s = \emptyset$ 
2  $W = \emptyset$ 
3 Add the first  $wsize$  edges to  $W$ 
4  $t = wsize$ 
5 Create a queue  $Q$ 
6 // uniformly sample a node from  $W$ 
7  $u = Uniform(V_W)$ 
8 for  $e_t$  in the graph stream  $S$  do
9   if  $u \notin V_s$  then add  $u$  to  $V_s$ 
10  if  $W.incident\_edges(u) \neq \emptyset$  then
11    Sample  $e$  from  $W.incident\_edges(u)$ 
12    Add  $e = (u, v)$  to  $E_s$ 
13    Remove  $e$  from  $W$ 
14    Add  $v$  to  $V_s$ 
15    Enqueue  $v$  onto  $Q$ 
16  else
17    if  $Q = \emptyset$  then  $u = Uniform(V_W)$ 
18    else  $u = Q.dequeue()$ 
19  Move the window  $W$ 
20  if  $|V_s| > n$  then
21    Retain  $[e] \subset E_s$  such that  $[e]$  has  $n$  nodes
22    Output  $G_s = (V_s, E_s)$ 
23   $t = t + 1$ 

```

a single pass, and adds deterministically the first m edges incident to n nodes of the stream to the sample. Once it achieves the target sample size, then for any streaming edge, it adds the incident nodes to the sample (probabilistically) by replacing other sampled nodes from the node set (selected uniformly at random). At each step, it will also add the edge to the sample if its two incident nodes are already in the sampled node set---producing a partial induction effect).

Now, we discuss the properties of PIES to illustrate its characteristics.

- (1) *PIES is a two-phase sampling method.* A two-phase sampling method is a method in which an initial sample of units is selected from the population (e.g., the graph stream), and then a second sample is selected as a subsample of the first. PIES can be analyzed as a two-phase sampling method. The first phase samples edges (i.e., edge sampling) from the graph stream with probability $p_e = \frac{m}{t}$ if the edge is incident to at least one node that does not belong to the reservoir (here m is the number of initial edges in the reservoir). Also, an edge is sampled with probability $p_e = 1$ if the edge is incident to two nodes that belong to the reservoir. After that, the second phase samples a subgraph uniformly (i.e., node sampling) to maintain only n nodes in the reservoir (i.e., all nodes in the reservoir are equally likely to be sampled).
- (2) *PIES has a selection bias to high degree nodes.* PIES is biased to high degree nodes due to its first phase that relies on edge sampling. Naturally, edge sampling is biased towards high degree nodes since they tend to have more edges compared to lower degree nodes.
- (3) *PIES samples an induced subgraph uniformly from the sub-sampled edge stream $E'(t)$ at any time t in the stream.* At any time t in the graph stream E , PIES subsamples $E(t)$ to $E'(t)$ (such that $|E'(t)| \leq |E(t)|$). Then, PIES samples a uniform induced subgraph from $E'(t)$, such that

all nodes in $E'(t)$ have an equal chance to be selected. Now, that we analyzed PIES, it can be easily adapted depending on a specific choice of the network sampling goal.

ALGORITHM 5: PIES(Sample Size n , Stream S)

Input : Sample Size n , Graph Stream S
Output : Sampled Subgraph $G_s = (V_s, E_s)$

```

1  $V_s = \emptyset, E_s = \emptyset$ 
2  $t = 1$ 
3 while graph is streaming do
4    $(u, v) = e_t$ 
5   if  $|V_s| < n$  then
6     if  $u \notin V_s$  then  $V_s = V_s \cup \{u\}$ 
7     if  $v \notin V_s$  then  $V_s = V_s \cup \{v\}$ 
8      $E_s = E_s \cup \{e_t\}$ 
9      $m = |E_s|$ 
10  else
11     $p_e = \frac{m}{t}$ 
12    draw  $r$  from continuous Uniform(0,1)
13    if  $r \leq p_e$  then
14      draw  $i$  and  $j$  from discrete Uniform[1,  $|V_s|$ ]
15      if  $u \notin V_s$  then  $V_s = V_s \cup \{u\}$ , drop node  $V_s[i]$  with all its incident edges
16      if  $v \notin V_s$  then  $V_s = V_s \cup \{v\}$ , drop node  $V_s[j]$  with all its incident edges
17    if  $u \in V_s$  AND  $v \in V_s$  then  $E_s = E_s \cup \{e_t\}$ 
18   $t = t + 1$ 

```

6.1. Experimental Evaluation

In this section, we present results of sampling from streaming graphs presented as an arbitrarily ordered sequence of edges. The experimental setup is similar to what we used in section 5.3. We compare the performance of our proposed algorithm PIES to the proposed streaming implementation of node (NS), edge (ES), and breadth-first search sampling (BFS) methods. Note that we implement breadth first search using a sliding window of 100 edges.

Similar to the experiments in section 5.3, we use ten different runs. To assess algorithm variation based on the edge sequence ordering, we randomly permute the edges in each run (while ensuring that all sampling methods use the same sequential order). Note that we compare the methods on a large social network from LIVEJOURNAL [SNAP] with 4 million nodes and 68 million edges (included only at the 20% sample size).

Distance metrics. Figures 5(a)--5(d) show the average KS statistic for degree, path length, clustering coefficient, and k-core distributions as an average over six datasets (as in section 5.3). PIES outperforms all other methods for the degree distribution statistic. NS performs almost as good as PIES for path length, clustering coefficient, and k-core distributions. As we explained in section 6, PIES is biased to high degree nodes (due to its first phase) compared to NS. Both BFS and ES performs the worst among the four methods. This shows that the limited observability of the graph structure using a window of 100 edges does not facilitate effective breadth-first search. While increasing the window size may help improve the performance of BFS, we did not explore this as our focus was primarily on space-efficient sampling. Similar to the results of KS statistic, Figures 5(e)--5(h) show the skew divergence statistic.

Finally, Figures 5(i)--5(j) show the L1 and L2 distance for eigenvalues and network values respectively. PIES outperforms all other methods. However, even though PIES performs the best,

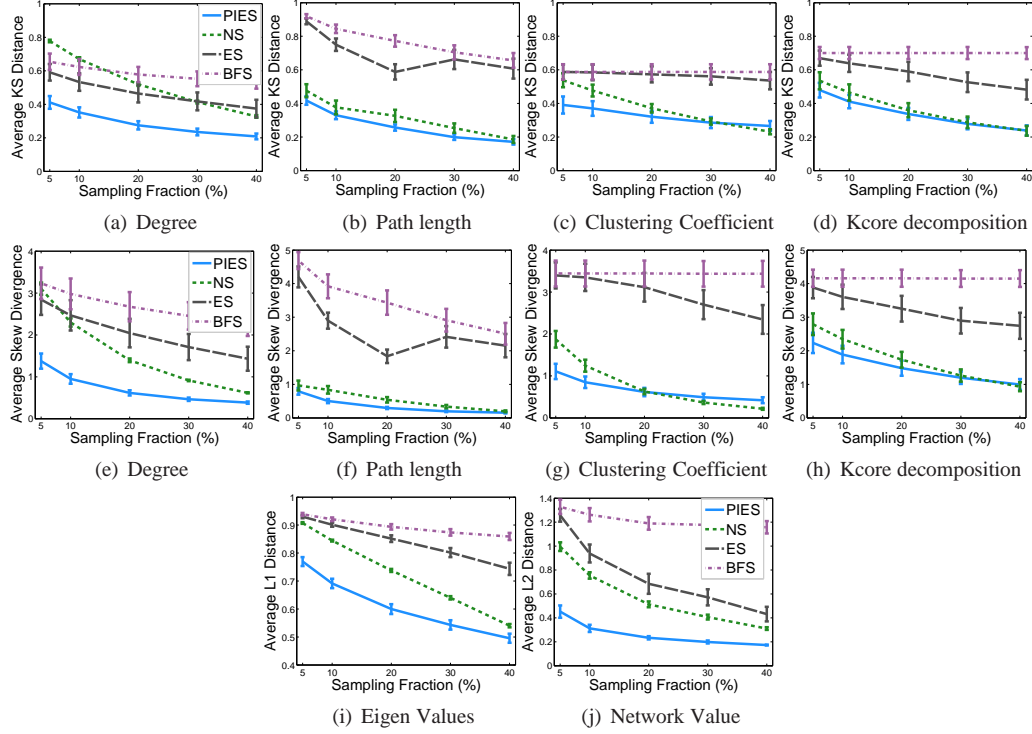


Fig. 5: (a-d) Average KS distance, (e-h) average skew divergence, and (i-j) average L1 and L2 distance respectively, across 6 datasets.

the distance is almost 50% for the eigenvalues. This implies PIES is not suitable for capturing the eigenvalues of the graph.

Distributions. We plot the distributions of the six network statistics at the 20% sample size. Figures 17, 19, 18, 20, 21, 22, and 23 (Appendix A) show the plots for all the distributions across the seven datasets.

Degree Distribution. We observe across the six datasets, PIES outperforms the other methods for FACEBOOK, TWITTER, EMAIL-UNIV, FLICKR, and LIVEJOURNAL. However, PIES only performs slightly better than NS for HEPH and CONDMAT. This behavior appears to be related to the specific properties of the network datasets themselves. HEPH and CONDMAT are more clustered and dense compared to other graphs used in the evaluation. We will discuss the behavior of the sampling methods for dense versus sparse graphs later in this section.

Path length Distribution. PIES preserves the path length distribution of FACEBOOK, TWITTER, EMAIL-UNIV, FLICKR, and LIVEJOURNAL, however, it overestimates the shortest path for HEPH and CONDMAT.

Clustering Distribution. PIES generally underestimates the clustering coefficient in the graph by missing some of the clustering surrounding the sampled nodes. This behavior is more clear in HEPH and CONDMAT since they are more clustered initially.

K-Core Distribution. Similar to the previous statistics, PIES outperforms the other methods for FACEBOOK, TWITTER, and LIVEJOURNAL. For HEPH and CONDMAT, PIES performs almost as good as NS. In addition to the distribution of the core sizes, we compared the *max-core*

number in the sampled subgraphs to their real counterparts for the 20% sample size (Table IV).

Eigenvalues. While PIES captures the eigenvalues better than ES and BFS, its eigenvalues are orders of magnitude smaller than the real graph's eigenvalues. This implies that none of the streaming algorithms captures the eigenvalues (compared to ES-i in section 5).

Network values. PIES accurately estimates the network values of most of the graphs compared to other methods.

Table IV: Comparison of *max-core-number* for the 20% sample size for PIES, NS, ES, BFS versus Real value of G

Graph	Real max core no.	PIES	NS	ES	BFS
HEPPH	30	8*	8*	2	1
CONDMAT	25	7*	7*	2	1
TWITTER	18	7*	4	3	1
FACEBOOK	16	6*	4	2	1
FLICKR	406	166*	81	19	1
LIVEJOURNAL	372	117*	82	5	1
EMAIL-UNIV	47	22*	15	3	1

Analysis of dense versus sparse graphs. Further to the discussion of the distributional results, we note that PIES is more accurate for sparse, less clustered graphs. To illustrate this, we report the performance of the stream sampling methods for each network in Figure 6, sorted from left to right in ascending order by clustering coefficient and density. Note that the bars represent the KS statistic (averaged over degree, path length, clustering, and k-core) for the 20% sample size. Clearly, the KS statistic for all methods increases as the graph becomes more dense and clustered. PIES maintains a KS distance of approximately $\leq 24\%$ for five out of seven networks. These results indicate that PIES performs better in networks that are generally sparse and less clustered. This interesting result shows that PIES will be more suitable to sample rapidly changing graph streams that have lower density and less clustering---which is likely to be the case for many large-scale dynamic communication and activity networks.

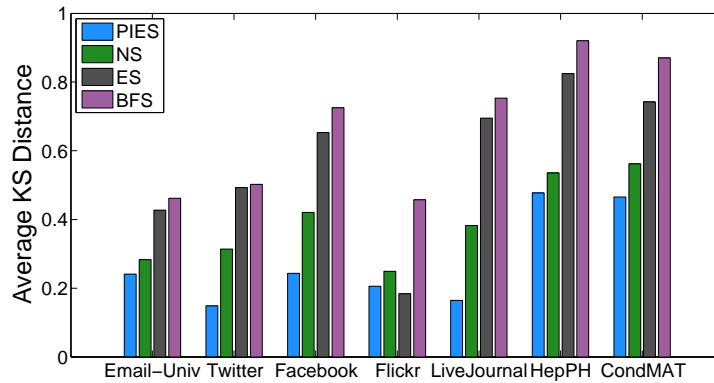


Fig. 6: Average KS Statistics for different networks (sorted in increasing order of clustering/density from left to right).

Moreover, we analyzed the number of isolated nodes for both NS and PIES in Table V. Since both NS and PIES sample nodes independently, it is expected that their sampled subgraph contains some nodes with zero degree (*i.e.*, isolated nodes). This implies that PIES carries isolated nodes in the reservoir as the graph streams by. From the process of PIES, we know each time a new edge is sampled from the stream, its incident nodes replace randomly selected nodes from the reservoir. This random replacement policy could replace high degree nodes while other isolated nodes remain in the reservoir. With this observation in mind, we propose a modification for PIES such that a newly added node replaces the node with minimum degree which has stayed in the reservoir the longest amount of time without acquiring more edges. This strategy favors retaining high degree nodes over isolated and/or low degree nodes in the sample. We show the results of this modification in Tables VI and VII which compare the KS distance, and L1/L2 distances respectively for each dataset (average over the two reasonable sample sizes 20% and 30%). Note that we refer to the modification of PIES as "PIES (MIN)". Clearly, modifying PIES in this manner (*i.e.*, PIES (MIN)) achieved better results for dense graphs such as HEPH and CONDMAT.

Table V: Average probability of isolated nodes at the 20% sample size for NS and PIES sampling methods.

Graph	PIES	NS
HEPPH	0.046	0.15
CONDMAT	0.14	0.36
TWITTER	0.15	0.51
FACEBOOK	0.13	0.43
FLICKR	0.14	0.56
LIVEJOURNAL	0.06	0.36
EMAIL-UNIV	0.07	0.51

Summary---We summarize the main empirical conclusions in this section:

- (1) *Sampled subgraphs collected and constructed by PIES accurately preserve many network statistics (e.g., degree, path length, and k-core).*
- (2) *PIES produces better samples when the graph is sparse/less clustered (e.g., TWITTER and LIVEJOURNAL).*
- (3) *We showed how PIES can be adapted to reduce the number of isolated nodes in the sample "PIES(MIN)".*
- (4) *PIES(MIN) can preserve the properties of dense graphs as well as certain statistics (e.g., eigenvalues) that are hard to be preserved by PIES.*
- (5) *The results show that the structure of the sampled subgraph G_s depends on the manner in which the topology of the graph G , the characteristics of η (e.g., degree distribution), and the nature of the sampling method interact. In future work, we aim to study how to adapt the sample given prior knowledge of the graph properties.*

7. PRACTICAL APPLICATIONS OF NETWORK SAMPLING

In Section 2, we discussed how network sampling arises in many different applications (*e.g.*, social science, data mining). Most research in network sampling has focused on how to collect a sample that closely match *topological* properties of the network [Leskovec and Faloutsos 2006; Maiya and Berger-Wolf 2011]. However, since the topological properties are never entirely preserved, it is also important to study how the sampling process impacts the investigation of applications overlaid on the networks. One such study recently investigated the impact of sampling methods on the discovery of the information diffusion process [De Choudhury et al. 2010]. The study

Table VI: Average KS Distance for PIES, PIES (MIN), NS, ES, and BFS stream sampling methods.

Data		PIES	PIES (MIN)	NS	ES	BFS
EMAIL-UNIV	<i>Deg</i>	0.2348	0.5803	0.4547	0.2186	0.3724
	<i>PL</i>	0.204	0.5621	0.19	0.6989	0.5114
	<i>Clust</i>	0.1108	0.4728	0.188	0.3302	0.3473
	<i>KCore</i>	0.2819	0.5821	0.1985	0.3219	0.5759
		0.2079*	0.5493	0.2578	0.3924	0.4518
TWITTER	<i>Deg</i>	0.1521	0.2598	0.4667	0.3052	0.4194
	<i>PL</i>	0.0528	0.3941	0.1243	0.617	0.4811
	<i>Clust</i>	0.2462	0.2269	0.346	0.4673	0.482
	<i>KCore</i>	0.1001	0.2886	0.2271	0.4393	0.5929
		0.1378*	0.2923	0.291	0.4572	0.4938
FACEBOOK	<i>Deg</i>	0.1848	0.2357	0.3804	0.4912	0.6917
	<i>PL</i>	0.2121	0.3171	0.4337	0.8762	0.9557
	<i>Clust</i>	0.2594	0.2314	0.3496	0.4975	0.5017
	<i>KCore</i>	0.2375	0.2447	0.3569	0.661	0.7275
		0.2234*	0.2572	0.3802	0.6315	0.7192
FLICKR	<i>Deg</i>	0.1503	0.399	0.514	0.0924	0.2706
	<i>PL</i>	0.2845	0.4936	0.0789	0.1487	0.6763
	<i>Clust</i>	0.1426	0.3754	0.2404	0.3156	0.3931
	<i>KCore</i>	0.1654	0.4289	0.0595	0.1295	0.4541
		0.1857	0.4242	0.2232	0.1716*	0.4485
HEPPH	<i>Deg</i>	0.4103	0.1304	0.483	0.8585	0.8923
	<i>PL</i>	0.306	0.1959	0.431	0.749	0.8676
	<i>Clust</i>	0.4636	0.0393	0.3441	0.9156	0.9171
	<i>KCore</i>	0.592	0.1674	0.6233	0.9402	0.9592
		0.443	0.1332*	0.4704	0.8658	0.909
CONDMAT	<i>Deg</i>	0.4042	0.1259	0.5006	0.6787	0.7471
	<i>PL</i>	0.2944	0.2758	0.5211	0.6981	0.9205
	<i>Clust</i>	0.5927	0.3285	0.5341	0.878	0.8853
	<i>KCore</i>	0.4692	0.1512	0.4955	0.858	0.8909
		0.4401	0.2203*	0.5128	0.7782	0.8609
Average for all Datasets		0.2730*	0.3128	0.3559	0.5494	0.6472

shows that sampling methods which considers both topology and user context improves on other naive methods. In this section, we consider the impact of sampling on relational learning. Network sampling is a core part of relational learning and it comes in many different problems. For example, learning models, evaluation of learning algorithms, learning ensembles of models, and active learning.

However, network sampling can produce samples with imbalance in class membership and bias in topological features (*e.g.*, path length, clustering) due to missing nodes/edges---thus the sampling process can significantly impact the accuracy of relational classification. This bias may result from the size of the sample, the sampling method, or both. While, most previous work in relational learning has focused on analyzing a single *input network* and research has considered how to further split the input network into training and testing networks for evaluation [Körner and Wrobel 2006; Macskassy and Provost 2007; Neville et al. 2009], the fact that the input network is often itself *sam-*

Table VII: Average L1/L2 Distance for PIES, PIES (MIN), NS, ES, and BFS stream sampling.

Data		PIES	PIES (MIN)	NS	ES	BFS
EMAIL-UNIV	<i>EigenVal</i>	0.4487	0.074*	0.7018	0.7588	0.838
	<i>NetVal</i>	0.199	0.0201*	0.5785	0.3799	1.007
TWITTER	<i>EigenVal</i>	0.4981	0.1851*	0.6411	0.7217	0.7964
	<i>NetVal</i>	0.1431	0.043*	0.3108	0.4271	0.8385
FACEBOOK (NO)	<i>EigenVal</i>	0.591	0.1143*	0.6771	0.8417	0.9018
	<i>NetVal</i>	0.306	0.0617*	0.5383	1.0984	1.5027
FLICKR	<i>EigenVal</i>	0.5503	0.0049*	0.7227	0.8491	0.9298
	<i>NetVal</i>	0.1657	0.0005*	0.5626	0.0574	1.5193
HEPPH	<i>EigenVal</i>	0.7083	0.2825*	0.7232	0.9373	0.95
	<i>NetVal</i>	0.3	0.1817*	0.3198	1.0821	1.2477
CONDMAT	<i>EigenVal</i>	0.6278	0.1475*	0.6843	0.8507	0.8875
	<i>NetVal</i>	0.2254	0.06*	0.3235	0.7514	0.9853

pled from an unknown target network has largely been ignored. There has been little focus on *how* the construction of the input networks may impact the evaluation of relational algorithms.

In this section, we study the question of how the choice of the sampling method can impact *parameter estimation* and *performance evaluation* of relational classification algorithms. We aim to evaluate the impact of network sampling on relational classification using two different goals:

- (1) *Parameter estimation*: we study the impact of network sampling on the estimation of class priors, *i.e.*, probability of class labels, goal 3.
- (2) *Performance evaluation*: we study the impact of network sampling on the estimation of classification accuracy of relational learners, *i.e.*, goal 2.

Case Study: Relational Classification

Conventional classification algorithms focus on the problem of identifying the unknown class (*e.g.*, group) to which an entity (*e.g.*, person) belongs. Classification models are learned from a training set of (disjoint) entities, which are assumed to be independent and identically distributed (*i.i.d.*) and drawn from the underlying population of instances. However, relational classification problems differs from this conventional view in that entities violate the *i.i.d.* assumption. In relational data, entities (*e.g.*, users in social networks) can exhibit complex dependencies. For example, friends often share similar interests (*e.g.*, political views).

Recently, there have been a great deal of research in relational learning and classification. For example, [Friedman et al. 1999] and [Taskar et al. 2001] outline probabilistic relational learning algorithms that search the space for relational attributes and structures of neighbors to improve the classification accuracy. Further, Macskassy proposed a simple relational neighbor classifier (weighted-vote relational neighbor wvRN) that requires no learning and iteratively classifying the entities of a relational network based only on the relational structure [Macskassy and Provost 2007]. Macskassy showed that wvRN often performs competitively to other relational learning algorithms.

Impact of sampling on parameter estimation. Let a be the node attribute representing the class label of any node $v_i \in V$ in graph G . We denote $\mathcal{C} = \{c_1, c_2, \dots\}$ as the set of possible class labels, where c_l is the class label of node v_i (*i.e.*, $a(v_i) = c_l$).

We study the impact of network sampling on the estimation of of class priors in G (*i.e.*, the distribution of class labels), using the following procedure:

- (1) Choose a set of nodes S from V using a sampling algorithm σ .
- (2) For each node $v_i \in S$, observe v_i 's class label.

(3) Estimate the class label distribution \hat{p}_{c_l} from S , using the following equation,

$$\hat{p}_{c_l} = \frac{1}{|S|} \sum_{v_i \in S} 1_{(a(v_i)=c_l)}$$

In our experiments, we consider four real networks: two citation networks CORA with 2708 nodes and CITESEER with 3312 nodes [Sen et al. 2008], FACEBOOK collected from Facebook Purdue network with 7315 users with their political views [Xiang et al. 2010], and a single day snapshot of 1490 political blogs that shows the interactions between liberal and conservative blogs [Adamic and Glance 2005]. We sample a subset of the nodes using NS, ES, ES-i, and FFS, where the sample size is between 10% – 80% of the graph G . For each sample size, we take the average of ten different runs. Then, we compare the estimated class prior to the actual class prior in the full graph G using the average KS distance measure. As plotted in Figures 7(a)–7(d), node sampling (NS) estimates the class priors more accurately than other methods, however, we note that FFS produces a large bias in most of the graphs (at 10% sample size).

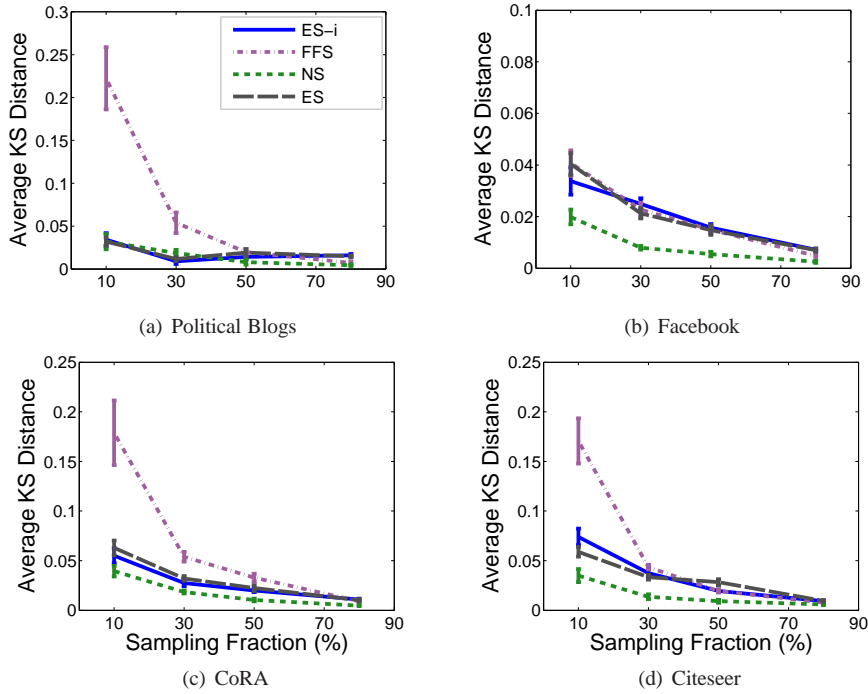


Fig. 7: Average KS distance of class priors for NS, ES, FFS, and ES-i.

Impact of sampling on classification accuracy. Let \mathcal{R} be a relational classifier which takes a graph G as input. The goal is to predict the class labels of nodes in G . Therefore, \mathcal{R} uses a proportion of nodes in graph G with known class labels as a *training set* to learn a model. Afterwards, \mathcal{R} is used to predict the label of the remaining (unlabeled) nodes in G (i.e., *test set*). Generally, the performance of \mathcal{R} can be evaluated based on the accuracy of the predicted class labels---we calculate the accuracy using area under the ROC curve (AUC) measure.

We study the impact of network sampling on the accuracy of relational classification using the following procedure:

- (1) Sample a subgraph G_s from G using a sampling algorithm σ .
- (2) Estimate the classification accuracy of a classifier \mathcal{R} on G_s : $\hat{auc} = \mathcal{R}(G_s)$

We compare the actual classification accuracy on G to the estimated classification accuracy on G_s . Formally, we compare $auc = \mathcal{R}(G)$ to $\hat{auc} = \mathcal{R}(G_s)$. So then, G_s is said to be representative to G , if $\hat{auc} \approx auc$.

In our experiments, we use the weighted-vote relational neighbor classifier (wvRN) as our base classifier \mathcal{R} [Macskassy and Provost 2007]. In wvRN, the class membership probability of a node v_i belonging to class c_l is defined as:

$$P(c_l|v_i) = \frac{1}{Z} \sum_{v_j \in \mathcal{N}(v_i)} w(v_i, v_j) * P(c_l|v_j)$$

where $\mathcal{N}(v_i)$ is the set of neighbors of node v_i , $w(v_i, v_j)$ is the weight of the edge $e_{ij} = (v_i, v_j)$, and $Z = \sum_{v_j \in \mathcal{N}(v_i)} w(v_i, v_j)$ is the normalization term.

We follow the common methodology used in [Macskassy and Provost 2007] to compute the classification accuracy. First, we vary the proportion of randomly selected labeled nodes from 10% – 80%; and we use 5-fold cross validation to compute the average AUC. Then, we repeat this procedure for both the graph G and the sample subgraph G_s . Note that AUC is calculated for the most prevalent class. Figures 8(a) -- 8(d) show the plots of AUC versus the sample size ($\phi = 10\% - 80\%$) with 10% labeled nodes. Similarly, Figures 9(a) -- 9(d) show the plots of AUC versus the proportion of labeled nodes, such that the AUC is an average of all sample sizes (10% – 80%). We observe that AUC of G is generally underestimated for sample sizes $< 30\%$ in the case of NS, ES, and FFS. However, generally ES-i performs better than other sampling methods and converges to the "True" AUC on G .

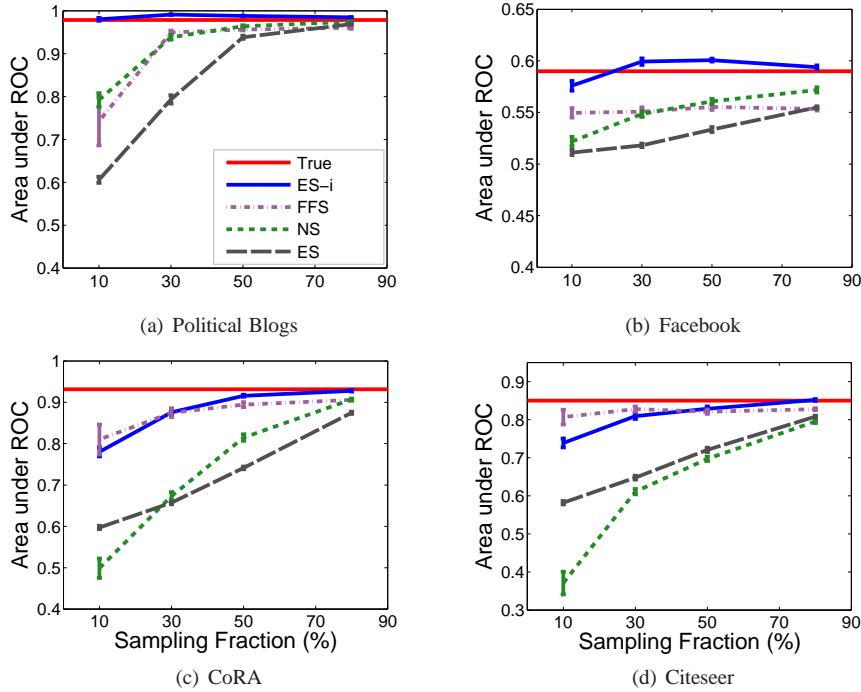


Fig. 8: Classification accuracy versus sampling fraction with 10% initially labeled nodes.

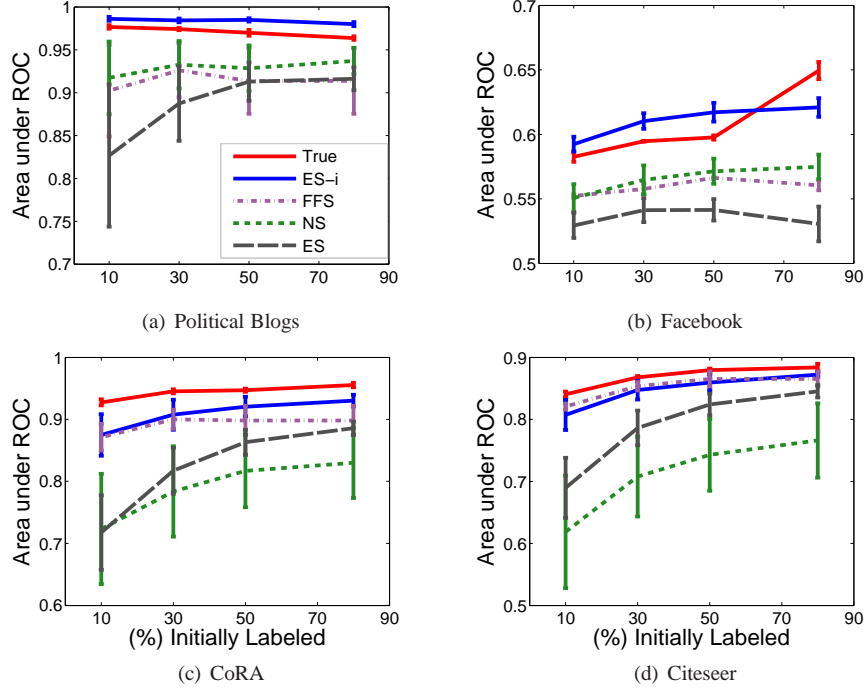


Fig. 9: Classification accuracy versus proportion of labeled nodes.

Summary. We conclude that many sampling methods fail to satisfy the different two goals (*i.e.*, parameter versus accuracy estimation). For example, while node sampling estimates the class prior better than other methods, it cannot estimate the classification accuracy. Edge sampling performs similar to node sampling. In addition, Forest fire sampling is generally non robust for estimating class priors (for $\phi \leq 30\%$). Generally, ES-i provides a good balance for satisfying the two goals with a little bias at the smaller sample sizes.

8. CONCLUSIONS AND FUTURE WORK

In this paper, we outlined a framework for the general problem of network sampling, by highlighting the different goals of study (from different research fields), population and units of interest, and classes of network sampling methods. This framework should facilitate the comparison of different sampling algorithms (strengths and weaknesses) relative to the particular goal of study. In addition, we proposed and discussed a spectrum of computational models for network sampling methods, going from the simple and least restrictive model of sampling from static graphs to the more realistic and restrictive model of sampling from streaming graphs. Within the context of the proposed spectrum, we designed a family of sampling methods based on the concept of graph induction that generalize across the full spectrum of computational models (from static to streaming), while efficiently preserving many of the topological properties of static and streaming graphs. Our experimental results indicate that our family of sampling methods more accurately preserves the underlying properties of the graph for both static and streaming graphs. Finally, we studied the impact of network sampling algorithms on the parameter estimation and performance evaluation of relational classification algorithms. Our results indicate our sampling method produces accurate estimates of classification accuracy. Concerning future work, we aim to investigate the performance of sub-linear time stream sampling methods for sampling a representative subgraph as well as extending to other task-based evaluations.

REFERENCES

- ADAMIC, L. AND GLANCE, N. 2005. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*. ACM, 36--43.
- AGGARWAL, C. 2006a. *Data streams: models and algorithms*. Springer.
- AGGARWAL, C., HAN, J., WANG, J., AND YU, P. 2003. A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*. VLDB Endowment, 81--92.
- AGGARWAL, C., LI, Y., YU, P., AND JIN, R. 2010a. On dense pattern mining in graph streams. *Proceedings of the VLDB Endowment* 3, 1-2, 975--984.
- AGGARWAL, C., ZHAO, Y., AND YU, P. 2010b. On clustering graph streams. In *Proceedings of the SIAM International Conference on Data Mining*.
- AGGARWAL, C., ZHAO, Y., AND YU, P. 2011. Outlier detection in graph streams. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. IEEE, 399--409.
- AGGARWAL, C. C. 2006b. On Biased Reservoir Sampling in the Presence of Stream Evolution. In *Proceedings of the 32nd International Conference on Very Large Data Bases*. VLDB '06. 607--618.
- AHMED, N., BERCHMANS, F., NEVILLE, J., AND KOMPELLA, R. 2010a. Time-based sampling of social network activity graphs. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*. ACM, 1--9.
- AHMED, N., NEVILLE, J., AND KOMPELLA, R. 2012. Space-efficient sampling from social activity streams. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. ACM, 53--60.
- AHMED, N. K., NEVILLE, J., AND KOMPELLA, R. 2010b. Reconsidering the foundations of network sampling. *WIN'10*.
- AHN, Y., HAN, S., KWAK, H., MOON, S., AND JEONG, H. 2007. Analysis of topological characteristics of huge online social networking services. In *WWW*. 835--844.
- AL HASAN, M. AND ZAKI, M. 2009. Output space sampling for graph patterns. *Proceedings of the VLDB Endowment* 2, 1, 730--741.
- ALLFACEBOOK.COM. http://allfacebook.com/facebook-marketing-infographic-engagement_b98277.
- ALVAREZ-HAMELIN, J., DALL'ASTA, L., BARRAT, A., AND VESPIGNANI, A. 2005. k-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases. *arXiv preprint cs/0511007*.
- AVRACHENKOV, K., RIBEIRO, B., AND TOWSLEY, D. 2010. Improving random walk estimation accuracy with uniform restarts. *Algorithms and Models for the Web-Graph*, 98--109.
- BABCOCK, B., BABU, S., DATAR, M., MOTWANI, R., AND WIDOM, J. 2002a. Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 1--16.
- BABCOCK, B., DATAR, M., AND MOTWANI, R. 2002b. Sampling from a moving window over streaming data. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 633--634.
- BACKSTROM, L. AND KLEINBERG, J. 2011. Network bucket testing. In *Proceedings of the 20th international conference on World wide web*. ACM, 615--624.
- BAKSHY, E., ROSENN, I., MARLOW, C., AND ADAMIC, L. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 519--528.
- BAR-YOSSEF, Z., KUMAR, R., AND SIVAKUMAR, D. 2002. Reductions in streaming algorithms with an application to counting triangles in graphs. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*. SODA '02. 623--632.
- BAYKAN, E., HENZINGER, M., KELLER, S., DE CASTELBERG, S., AND KINZLER, M. 2009. A comparison of techniques for sampling web pages. *Arxiv preprint arXiv:0902.1604*.
- BUCHSBAUM, A., GIANCARLO, R., AND WESTBROOK, J. 2003. On finding common neighborhoods in massive graphs. *Theoretical Computer Science* 299, 1, 707--718.
- BURIOL, L., FRAHLING, G., LEONARDI, S., MARCHETTI-SPACCAMELA, A., AND SOHLER, C. 2006. Counting triangles in data streams. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 253--262.
- CARMİ, S., HAVLIN, S., KIRKPATRICK, S., SHAVITT, Y., AND SHIR, E. 2007. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences* 104, 27, 11150--11154.
- CHAKRABARTI, D., ZHAN, Y., AND FALOUTSOS, C. 2004. R-mat: A recursive model for graph mining. *Computer Science Department*, 541.
- CHARIKAR, M., CHEN, K., AND FARACH-COLTON, M. 2002. Finding frequent items in data streams. *Automata, Languages and Programming*, 784--784.
- CHEN, L. AND WANG, C. 2010. Continuous subgraph pattern search over certain and uncertain graph streams. *Knowledge and Data Engineering, IEEE Transactions on* 22, 8, 1093--1109.

- CORMODE, G. AND MUTHUKRISHNAN, S. 2005. Space efficient mining of multigraph streams. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 271--282.
- DASGUPTA, A., KUMAR, R., AND SIVAKUMAR, D. 2012. Social sampling. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 235--243.
- DE CHOUDHURY, M., LIN, Y., SUNDARAM, H., CANDAN, K., XIE, L., AND KELLIHER, A. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. 34--41.
- DOMINGOS, P. AND HULTEN, G. 2000. Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 71--80.
- FAN, W. 2004a. Streamminer: a classifier ensemble-based engine to mine concept-drifting data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 1257--1260.
- FAN, W. 2004b. Systematic data selection to mine concept-drifting data streams. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 128--137.
- FAN, W., HUANG, Y., WANG, H., AND YU, P. 2004. Active mining of data streams. In *Proc. of the 4th SIAM International Conference on Data Mining*. 457--461.
- FRANK, O. 1977. Survey sampling in graphs. *Journal of Statistical Planning and Inference* 1, 3, 235--264.
- FRANK, O. 1980. Sampling and inference in a population graph. *International Statistical Review/Revue Internationale de Statistique*, 33--41.
- FRANK, O. 1981. A survey of statistical methods for graph analysis. *Sociological methodology* 12, 110--155.
- FRIEDMAN, N., GETOOR, L., KOLLER, D., AND PFEFFER, A. 1999. Learning probabilistic relational models. In *IJCAI*.
- GABER, M., ZASLAVSKY, A., AND KRISHNASWAMY, S. 2005. Mining data streams: a review. *ACM Sigmod Record* 34, 2.
- GAO, J., FAN, W., HAN, J., AND YU, P. 2007. A general framework for mining concept-drifting data streams with skewed distributions. *Proc. of SIAM ICDM*.
- GILE, K. AND HANDCOCK, M. 2010. Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology* 40, 1, 285--327.
- GJOKA, M., KURANT, M., BUTTS, C., AND MARKOPOULOU, A. 2010. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*. Ieee, 1--9.
- GJOKA, M., KURANT, M., BUTTS, C., AND MARKOPOULOU, A. 2011. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on* 29, 9, 1872--1892.
- GKANTSIDIS, C., MIHAIL, M., AND SABERI, A. 2004. Random walks in peer-to-peer networks. In *IEEE INFOCOM*.
- GLEICH, D. F. 2012. Graph of flickr photo-sharing social network crawled in may 2006.
- GOLAB, L. AND ÖZSU, M. 2003. Issues in data stream management. *ACM Sigmod Record* 32, 2, 5--14.
- GOODMAN, L. 1961. Snowball sampling. *The Annals of Mathematical Statistics* 32, 1, 148--170.
- GRANOVETTER, M. 1976. Network sampling: Some first steps. *American Journal of Sociology*, 1287--1303.
- GUHA, S., MEYERSON, A., MISHRA, N., MOTWANI, R., AND O'CALLAGHAN, L. 2003. Clustering data streams: Theory and practice. *Knowledge and Data Engineering, IEEE Transactions on* 15, 3, 515--528.
- HAEMERS, W. 1995. Interlacing eigenvalues and graphs. *Linear Algebra and its Applications* 226, 593--616.
- HANSEN, M. AND HURWITZ, W. 1943. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 333--362.
- HECKATHORN, D. 1997. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, 174--199.
- HENZINGER, M., HEYDON, A., MITZENMACHER, M., AND NAJORK, M. 2000. On near-uniform url sampling. *Computer Networks* 33, 1, 295--308.
- HENZINGER, M., RAGHAVAN, P., AND RAJAGOPALAN, S. 1999. Computing on data streams. In *External Memory Algorithms: Dimacs Workshop External Memory and Visualization, May 20-22, 1998*. Vol. 50. Amer Mathematical Society, 107.
- HUBLER, C., KRIEGEL, H.-P., BORWARDT, K. M., AND GHAHRAMANI, Z. 2008. Metropolis algorithms for representative subgraph sampling. In *ICDM*.
- HULTEN, G., SPENCER, L., AND DOMINGOS, P. 2001. Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 97--106.
- JIA, Y., HOBEROCK, J., GARLAND, M., AND HART, J. 2008. On the visualization of social and other scale-free networks. *Visualization and Computer Graphics, IEEE Transactions on* 14, 6, 1285--1292.
- KLEINBERG, J., KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999. The web as a graph: Measurements, models, and methods. *Computing and Combinatorics*, 1--17.
- KOLACZYK, E. 2009. Sampling and estimation in network graphs. *Statistical Analysis of Network Data*, 1--30.
- KÖRNER, C. AND WROBEL, S. 2006. Bias-free hypothesis evaluation in multirelational domains. In *PAKDD*. 668--672.

- KRISHNAMURTHY, V., FALOUTSOS, M., CHROBAK, M., CUI, J., LAO, L., AND PERCUS, A. 2007. Sampling large Internet topologies for simulation purposes. *Computer Networks* 51, 15, 4284--4302.
- KUMAR, R., NOVAK, J., AND TOMKINS, A. 2010. Structure and evolution of online social networks. *Link Mining: Models, Algorithms, and Applications*, 337--357.
- KURANT, M., MARKOPOULOU, A., AND THIRAN, P. 2011. Towards unbiased bfs sampling. *Selected Areas in Communications, IEEE Journal on* 29, 9, 1799--1809.
- LAKHINA, A., BYERS, J., CROVELLA, M., AND XIE, P. 2003. Sampling biases in ip topology measurements. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*. Vol. 1. IEEE, 332--341.
- LEE, L. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *AI and Statistics*.
- LEE, S., KIM, P., AND JEONG, H. 2006. Statistical properties of sampled networks. *Physical Review E* 73, 016102.
- LESKOVEC, J. AND FALOUTSOS, C. 2006. Sampling from large graphs. In *SIGKDD*. 631--636.
- LI, X., YU, P., LIU, B., AND NG, S. 2009. Positive unlabeled learning for data stream classification. *SDM, SIAM*.
- MACSKASSY, S. AND PROVOST, F. 2007. Classification in networked data: A toolkit and a univariate case study. *JMLR* 8, May, 935--983.
- MAIYA, A. S. AND BERGER-WOLF, T. Y. 2010. Sampling Community Structure. In *WWW*.
- MAIYA, A. S. AND BERGER-WOLF, T. Y. 2011. Benefits of bias: Towards better characterization of network sampling. In *SIGKDD*.
- MANKU, G. S. AND MOTWANI, R. 2002. Approximate Frequency Counts over Data Streams. In *Proceedings of the 28th International Conference on Very Large Data Bases. VLDB '02*. 346--357.
- MCGREGOR, A. 2009. Graph mining on streams. *Encyclopedia of Database Systems*, 1271--1275.
- MISLOVE, A., MARCON, M., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. 2007. Measurement and analysis of online social networks. In *ACM/USENIX IMC*.
- MUTHUKRISHNAN, S. 2005. *Data streams: Algorithms and applications*. Now Publishers Inc.
- NEVILLE, J., GALLAGHER, B., AND ELIASSI-RAD, T. 2009. Evaluating statistical tests for within-network classifiers of relational data. In *ICDM*.
- NEWMAN, M. 2002. Assortative mixing in networks. *Physical Review Letters* 89, 20.
- PAPAGELIS, M., DAS, G., AND KOUDAS, N. 2011. Sampling online social networks.
- RASTI, A., TORKJAZI, M., REJAIE, R., DUFFIELD, N., WILLINGER, W., AND STUTZBACH, D. 2009. Respondent-driven sampling for characterizing unstructured overlays. In *INFOCOM 2009, IEEE*. IEEE, 2701--2705.
- REDNER, S. 1998. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems* 4, 2, 131--134.
- RIBEIRO, B. AND TOWSLEY, D. 2010. Estimating and sampling graphs with multidimensional random walks. In *ACM SIGCOMM Internet Measurement Conference*.
- ROSSI, R. AND NEVILLE, J. 2012. Time-evolving relational classification and ensemble methods. *Advances in Knowledge Discovery and Data Mining* 7301, 1--13.
- SARMA, A. D., GOLLAPUDI, S., AND PANIGRAHY, R. 2008. Estimating PageRank on Graph Streams. In *PODS*.
- SEN, P., NAMATA, G., BILGIC, M., GETOOR, L., GALLIGHER, B., AND ELIASSI-RAD, T. 2008. Collective classification in network data. *AI magazine* 29, 3, 93.
- SESHADHRI, C., PINAR, A., AND KOLDA, T. 2011. An in-depth analysis of stochastic kronecker graphs. *arXiv preprint arXiv:1102.5046*.
- SNAP. Stanford large network dataset collection. <http://snap.stanford.edu/data/index.html>.
- STUMPF, M., WIUF, C., AND MAY, R. 2005. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences* 102, 12, 4221--4224.
- STUTZBACH, D., REJAIE, R., DUFFIELD, N., SEN, S., AND WILLINGER, W. 2006. On unbiased sampling for unstructured peer-to-peer networks. In *IMC*. 27--40.
- TASKAR, B., SEGAL, E., AND KOLLER, D. 2001. Probabilistic classification and clustering in relational data. In *IJCAI*. 870--878.
- TATBUL, N., ÇETINTEMEL, U., ZDONIK, S., CHERNIACK, M., AND STONEBRAKER, M. 2003. Load shedding in a data stream manager. In *Proceedings of the 29th international conference on Very large data bases-Volume 29. VLDB Endowment*, 309--320.
- VATTANI, A., CHAKRABARTI, D., AND GUREVICH, M. 2011. Preserving personalized pagerank in subgraphs. In *Proceedings of ICML*.
- VISWANATH, B., MISLOVE, A., CHA, M., AND GUMMADI, K. P. 2009. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*.
- VITTER, J. 1985. Random sampling with a reservoir. *ACM Trans. Math. Softw.* 11.

- WANG, H., FAN, W., YU, P., AND HAN, J. 2003. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 226--235.
- WANG, H., YU, P., AND HAN, J. 2005. Mining data streams. *Data Mining and Knowledge Discovery Handbook*, 777--792.
- WATTERS, J. AND BIERNACKI, P. 1989. Targeted sampling: options for the study of hidden populations. *Social Problems*, 416--430.
- WATTS, D. AND STROGATZ, S. 1998. The small world problem. *Collective Dynamics of Small-World Networks* 393, 440--442.
- WILSON, C., BOE, B., SALA, A., PUTTASWAMY, K. P., AND ZHAO, B. Y. 2009. User interactions in social networks and their implications. In *EuroSys*. 205--218.
- XIANG, R., NEVILLE, J., AND ROGATI, M. 2010. Modeling relationship strength in online social networks. In *WWW*.
- YE, S., LANG, J., AND WU, F. 2010. Crawling online social graphs. In *Web Conference (APWEB), 2010 12th International Asia-Pacific*. IEEE, 236--242.
- YOON, S., LEE, S., YOOK, S.-H., AND KIM, Y. 2007. Statistical properties of sampled networks by random walks. *Phys. Rev. E* 75, 4, 046114.
- ZHANG, J. 2010. A survey on streaming algorithms for massive graphs. *Managing and Mining Graph Data*, 393--420.

A. APPENDIX A

A.1. Distributions for Static Graphs (at 20% sample size)

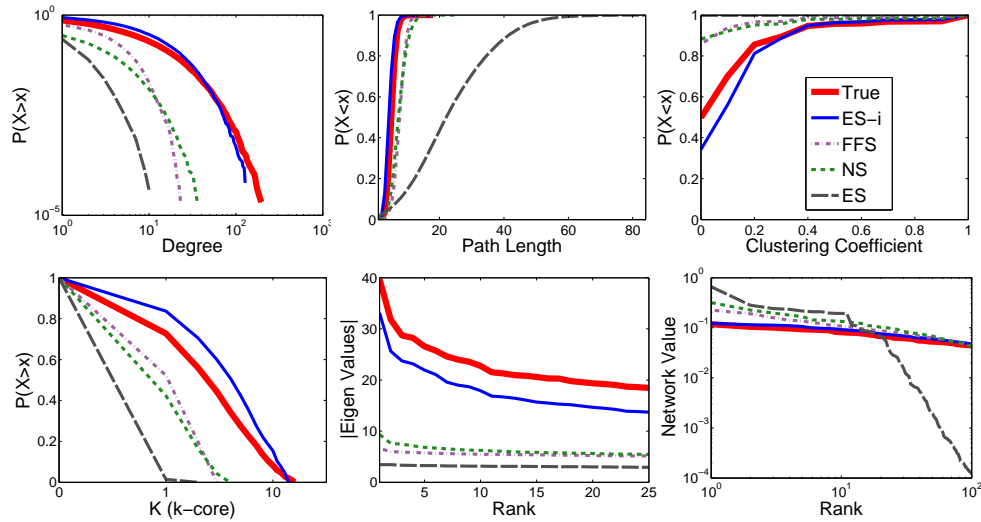


Fig. 10: FACEBOOK Graph

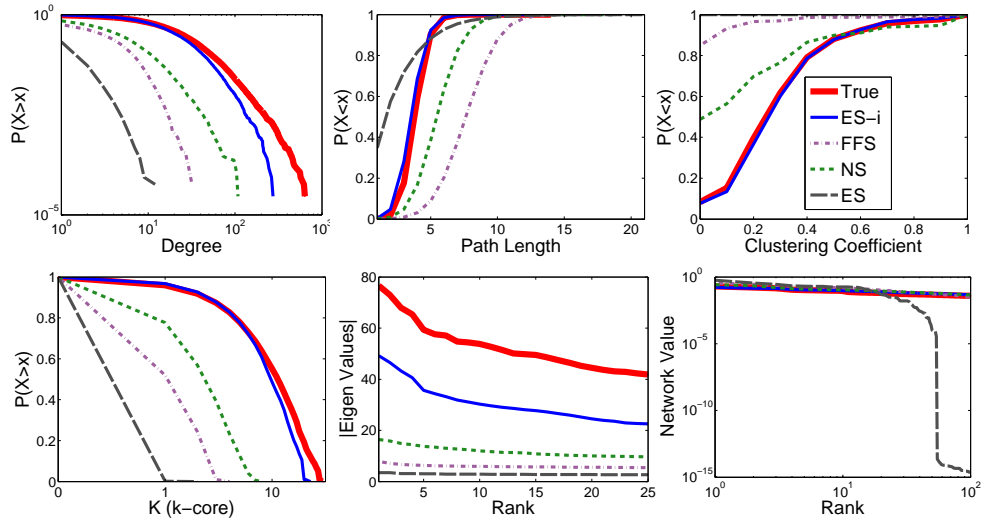


Fig. 11: HEPH Graph

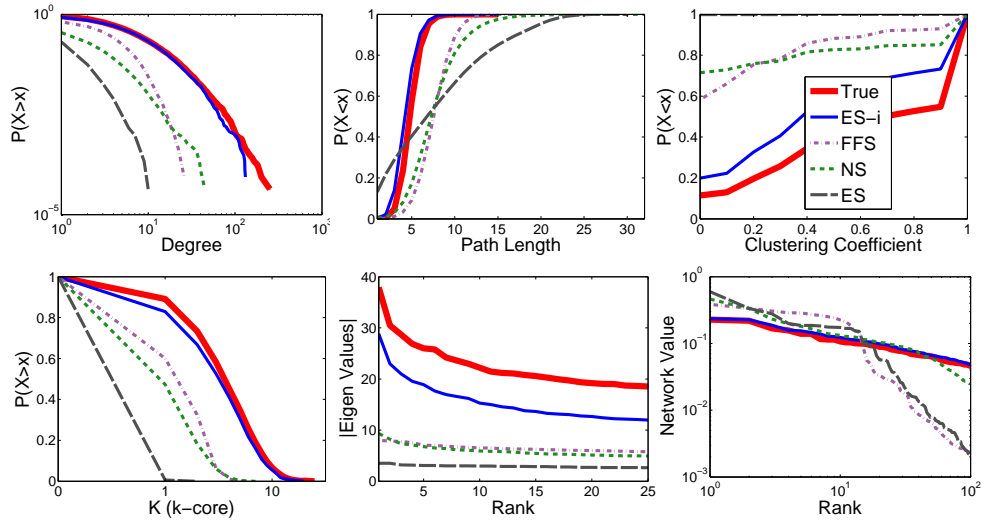


Fig. 12: CONDMAT Graph

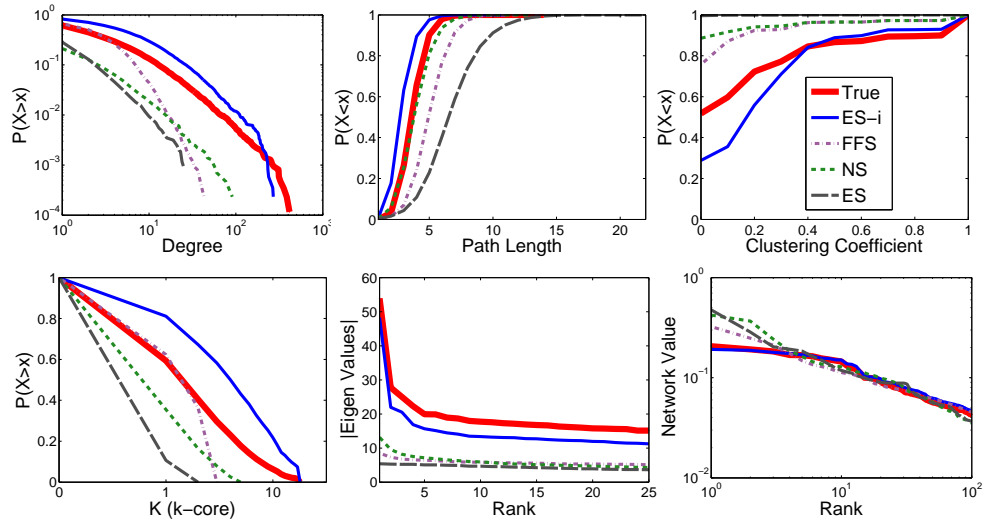


Fig. 13: TWITTER Graph

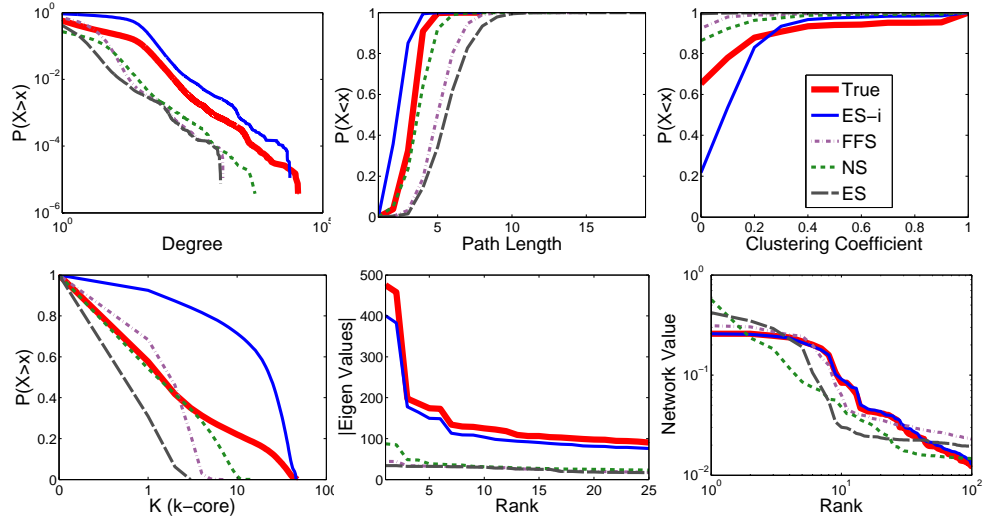


Fig. 14: EMAIL-UNIV Graph

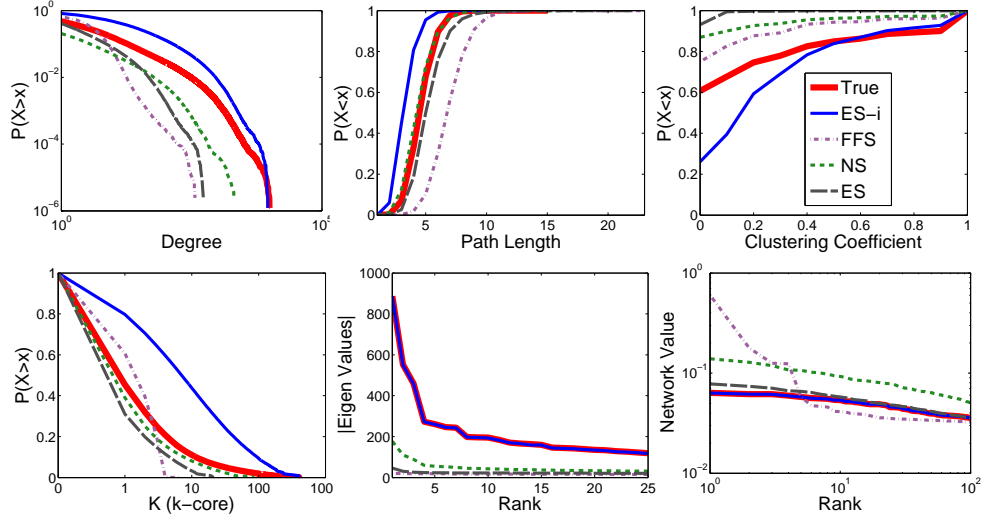


Fig. 15: FLICKR Graph

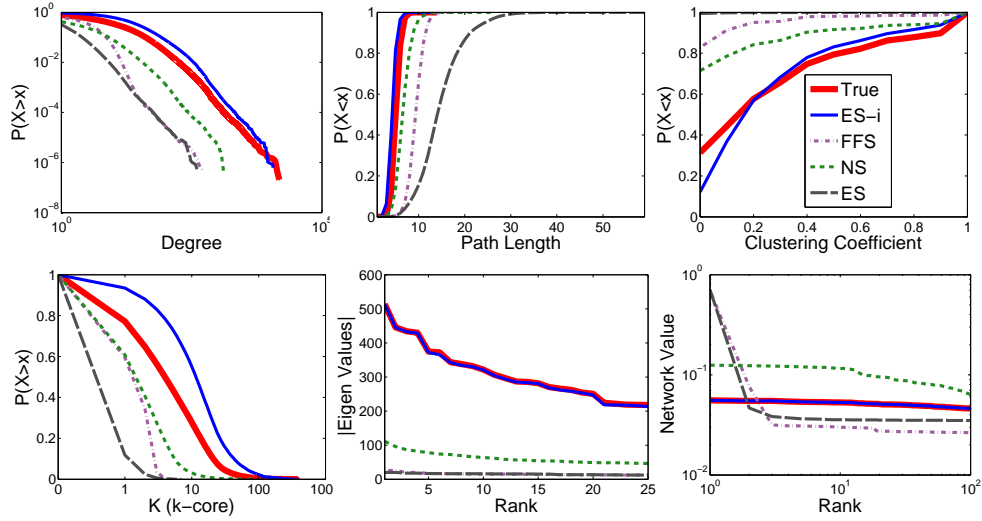


Fig. 16: LIVEJOURNAL Graph

A.2. Distributions for Streaming Graphs (at 20% sample size)

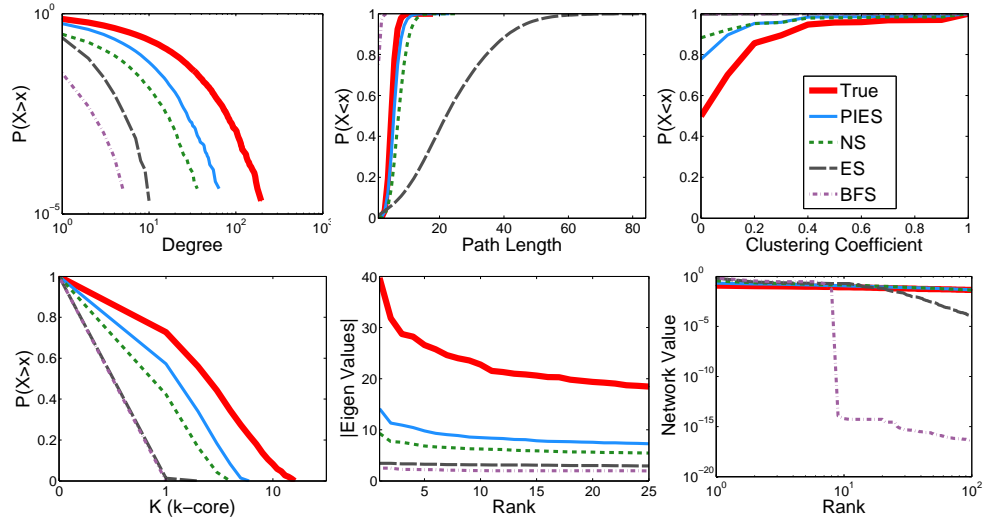


Fig. 17: FACEBOOK Graph

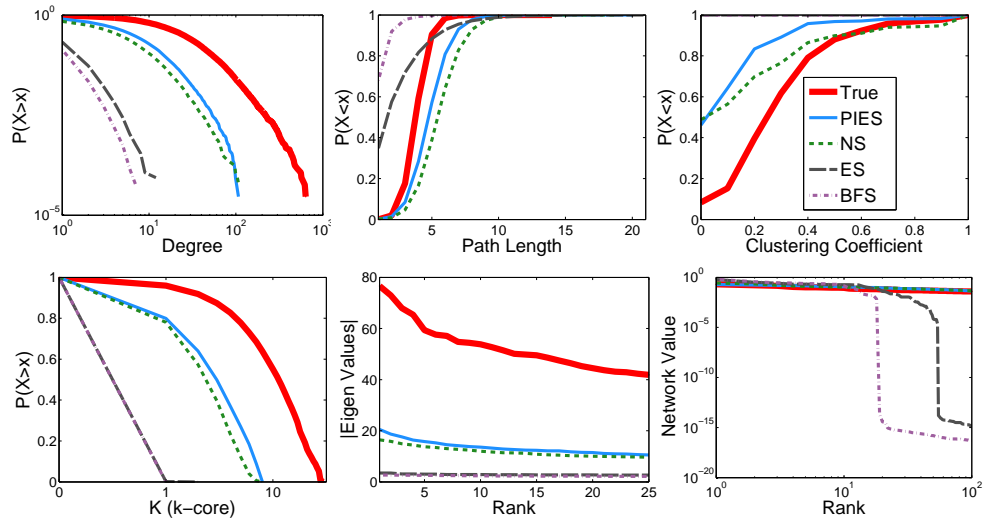


Fig. 18: HEPH Graph

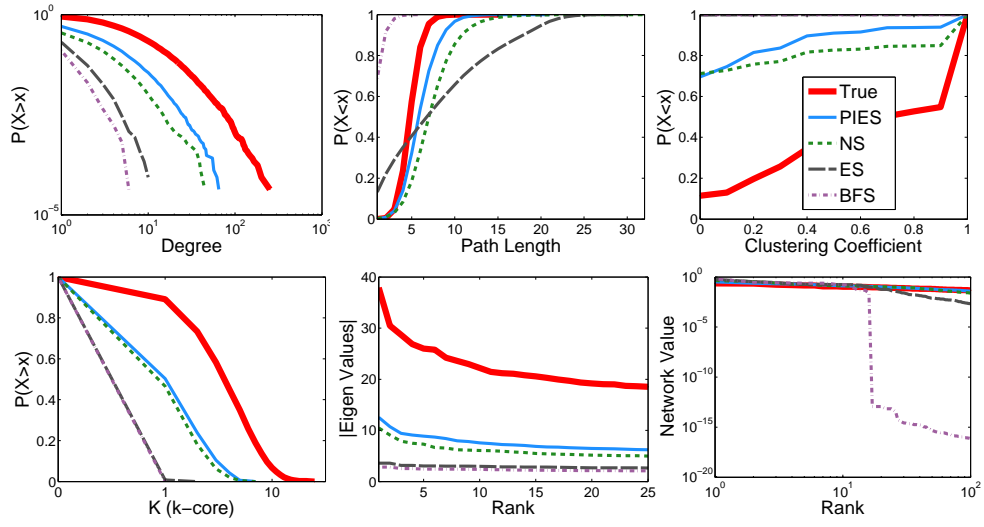


Fig. 19: CONDMAT Graph

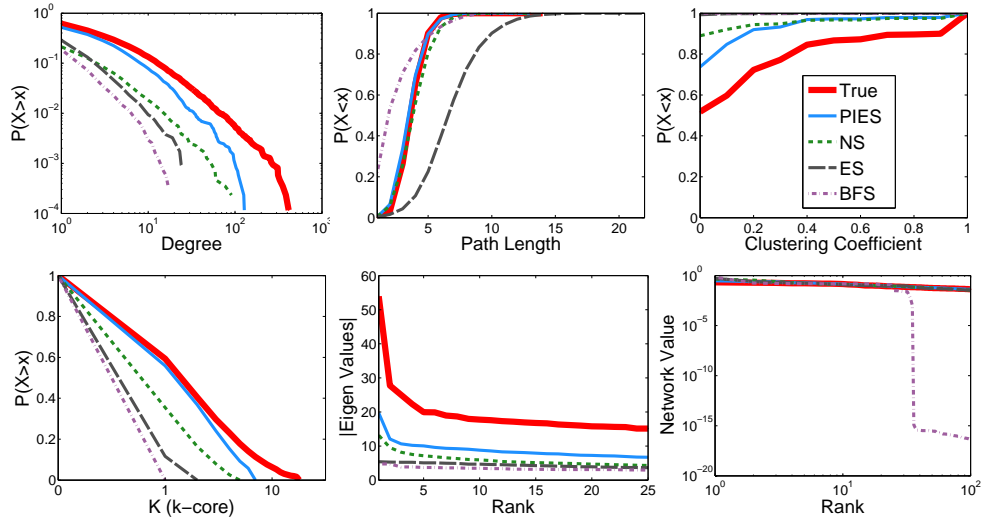


Fig. 20: TWITTER Graph

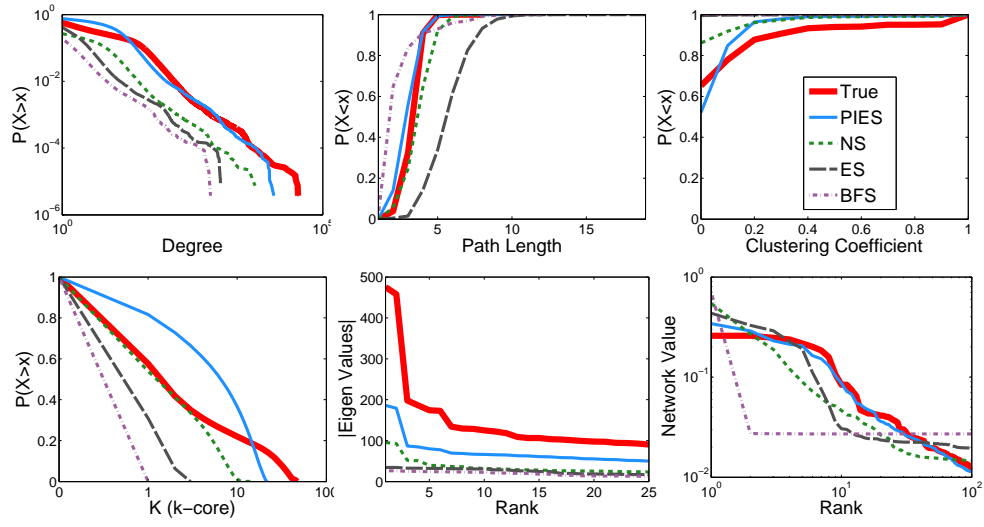


Fig. 21: EMAIL-UNIV Graph

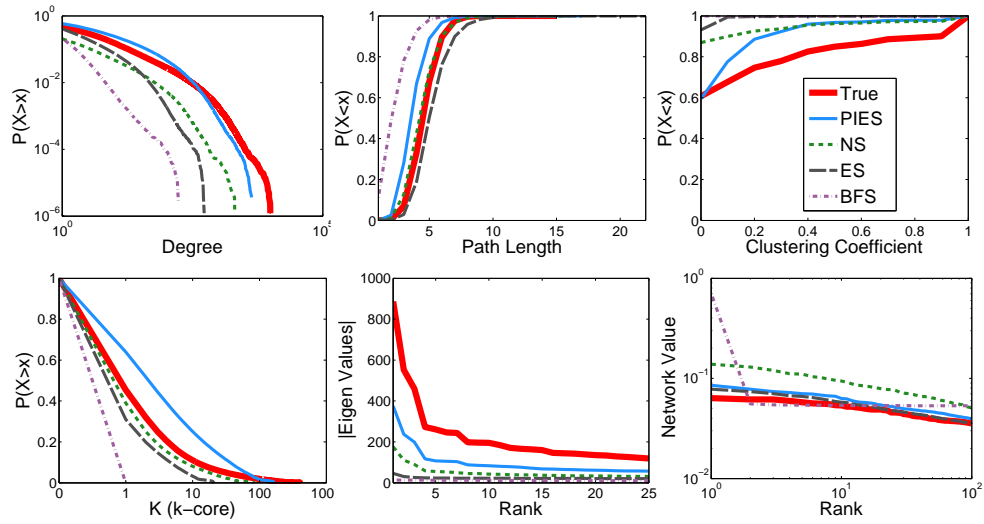


Fig. 22: FLICKR Graph

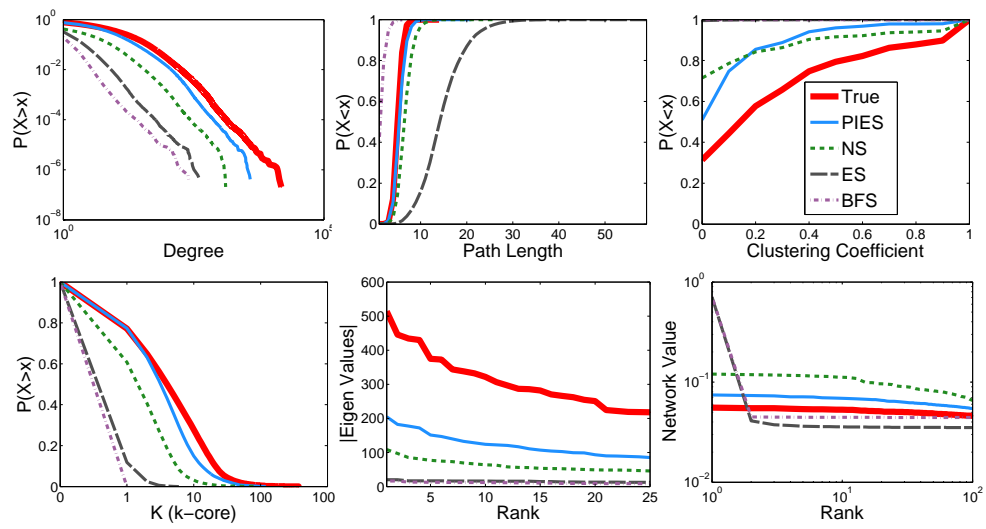


Fig. 23: LIVEJOURNAL Graph